METHODS FOR IDENTIFYING RISK OF BREAST CANCER AND TREATMENTS THEREOF

Related Patent Applications

[0001] This patent application claims the benefit of provisional patent application no. 60/429,136 filed November 25, 2002 and provisional patent application no. 60/490,234 filed July 24, 2003, having attorney docket number 524593004100 and 524593004101, respectively. Each of these provisional patent applications names Richard B. Roth *et al.* as inventors and is hereby incorporated herein by reference in its entirety, including all drawings and cited publications and documents.

Field of the Invention

[0002] The invention relates to genetic methods for identifying risk of breast cancer and treatments that specifically target the disease.

Background

[0003] Breast cancer is the third most common cancer, and the most common cancer in women, as well as a cause of disability, psychological trauma, and economic loss. Breast cancer is the second most common cause of cancer death in women in the United States, in particular for women between the ages of 15 and 54, and the leading cause of cancer-related death (Forbes, Seminars in Oncology, vol.24(1), Suppl 1, 1997: pp.S1-20-S1-35). Indirect effects of the disease also contribute to the mortality from breast cancer including consequences of advanced disease, such as metastases to the bone or brain. Complications arising from bone marrow suppression, radiation fibrosis and neutropenic sepsis, collateral effects from therapeutic interventions, such as surgery, radiation, chemotherapy, or bone marrow transplantation-also contribute to the morbidity and mortality from this disease.

[0004] While the pathogenesis of breast cancer is unclear, transformation of normal breast epithelium to a malignant phenotype may be the result of genetic factors, especially in women under thirty (Miki, *et al.*, Science, 266: 66-71 (1994)). However, it is likely that other, non-genetic factors also have a significant effect on the etiology of the disease. Regardless of its origin, breast cancer morbidity increases significantly if it is not detected early in its progression. Thus, considerable efforts have focused on the elucidation of early cellular events surrounding transformation in breast tissue. Such efforts have led to the identification of several potential breast cancer markers. For example, alleles of the *BRCA1* and *BRCA2* genes have been linked to hereditary and early-onset breast cancer (Wooster, *et al.*, Science, 265: 2088-2090 (1994)). However, *BRCA1* is limited as a cancer marker because *BRCA1*

mutations fail to account for the majority of breast cancers (Ford, *et al.*, British J. Cancer, 72: 805-812 (1995)). Similarly, the *BRCA2* gene, which has been linked to forms of hereditary breast cancer, accounts for only a small portion of total breast cancer cases.

<u>Summary</u>

[0005] It has been discovered that certain polymorphic variations in human genomic DNA are associated with the occurrence of breast cancer. In particular, polymorphic variants in loci containing *GP6, LAMA4, CHGB/C20orf154* (hereafter referred to as "*CHGB*"), *LOC338749* and *TTN/LOC351327* (hereafter referred to as "*TTN*") regions in human genomic DNA have been associated with risk of breast cancer.

[0006] Thus, featured herein are methods for identifying a subject at risk of breast cancer and/or a risk of breast cancer in a subject, which comprises detecting the presence or absence of one or more polymorphic variations associated with breast cancer in genomic regions described herein in a human nucleic acid sample. In an embodiment, two or more polymorphic variations are detected in two or more regions selected from the group consisting of *GP6, LAMA4, CHGB, LOC338749* and *TTN*. In certain embodiments, 3 or fewer, or 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 or fewer polymorphic variants are detected.

[0007] Also featured are nucleic acids that include one or more polymorphic variations associated with the occurrence of breast cancer, as well as polypeptides encoded by these nucleic acids. Further, provided is a method for identifying a subject at risk of breast cancer and then prescribing to the subject a breast cancer detection procedure, prevention procedure and/or a treatment procedure. In addition, provided are methods for identifying candidate therapeutic molecules for treating breast cancer and related disorders, as well as methods for treating breast cancer in a subject by diagnosing breast cancer in the subject and treating the subject with a suitable treatment, such as administering a therapeutic molecule.

[0008] Also provided are compositions comprising a breast cancer cell and/or *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid with a RNAi, siRNA, antisense DNA or RNA, or ribozyme nucleic acid designed from a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence. In an embodiment, the nucleic acid is designed from a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence that includes one or more breast cancer associated polymorphic variations, and in some instances, specifically interacts with such a nucleotide sequence. Further, provided are arrays of nucleic acids bound to a solid surface, in which one or more nucleic acid molecules of the array have a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence, or a fragment or substantially identical nucleic acid thereof, or a complementary nucleic acid of the foregoing. Featured also are compositions comprising a breast

cancer cell and/or a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide, with an antibody that specifically binds to the polypeptide. In an embodiment, the antibody specifically binds to an epitope in the polypeptide that includes a non-synonymous amino acid modification associated with breast cancer (e.g., results in an amino acid substitution in the encoded polypeptide associated with breast cancer). In certain embodiments, the antibody specifically binds to an epitope that comprises a a lysine at amino acid position 237 of SEQ ID NO: 12, a proline at amino acid position 413 of SEQ ID NO: 16 or a glutamine at amino acid position 63 of SEQ ID NO: 16.

Brief Description of the Figures

[0009] Figures 1A-1Y show a genomic nucleotide sequence for an *GP6* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 1. The following nucleotide representations are used throughout: "A" or "a" is adenosine, adenine, or adenylic acid; "C" or "c" is cytidine, cytosine, or cytidylic acid; "G" or "g" is guanosine, guanine, or guanylic acid; "T" or "t" is thymidine, thymine, or thymidylic acid; and "I" or "i" is inosine, hypoxanthine, or inosinic acid. Exons are indicated in italicized lower case type, introns are depicted in normal text lower case type, and polymorphic sites are depicted in bold upper case type. SNPs are designated by the following convention: "R" represents A or G, "M" represents A or C; "W" represents A or T; "Y" represents C or T; "S" represents C or G; "K" represents G or T; "V" represents A, C or G; "H" represents A, C, or T; "D" represents A, G, or T; "B" represents C, G, or T; and "N" represents A, G, C, or T.

[0010] Figures 2A-2Y show a genomic nucleotide sequence of a *LAMA4* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 2.

[0011] Figures 3A-3X show a genomic nucleotide sequence of a *CHGB* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 3.

[0012] Figures 4A-4Y show a genomic nucleotide sequence of a *LOC338749* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 4.

[0013] Figures 5A-5Z show a genomic nucleotide sequence of a *TTN* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 5.

[0014] Figures 6A-6B show three coding nucleotide sequences (cDNA) for *GP6*. The nucleotide sequence are set forth as SEQ ID NO: 6-8.

[0015] Figures 7A-7B show a coding nucleotide sequence (cDNA) for *LAMA4*. The nucleotide sequence is set forth in SEQ ID NO: 9.

[0016] Figure 8 shows a coding nucleotide sequence (cDNA) for *CHGB*. The nucleotide sequence is set forth in SEQ ID NO: 10.

[0017] Figure 9 shows a coding nucleotide sequence (cDNA) for *TTN*. The nucleotide sequence is set forth in SEQ ID NO: 11.

[0018] Figures 10A-10C show three *GP6* polypeptide amino acid sequences, which are set forth as SEQ ID NO: 12-14.

[0019] Figures 11A-11B show an amino acid sequence for a *LAMA4* polypeptide, which is set forth in SEQ ID NO: 15.

[0020] Figure 12 shows an amino acid sequence for a *CHGB* polypeptide, which is set forth in SEQ ID NO: 16.

[0021] Figure 13 shows an amino acid sequence for a *TTN* polypeptide, which is set forth in SEQ ID NO: 17.

[0022] Figures 14-18 show proximal SNPs in *GP6, LAMA4, CHGB, LOC338749* and *TTN* loci in genomic DNA. The position of each SNP on the chromosome is shown on the x-axis and the y-axis provides the negative logarithm of the p-value comparing the estimated allele to that of the control group. Also shown in the figure are exons and introns of the genes in the approximate chromosomal positions. The figure indicates that polymorphic variants associated with breast cancer are in linkage disequilibrium in the following regions: the region spanning positions 185-8377 in SEQ ID NO: 1; the region spanning positions 506-95220 in SEQ ID NO: 2; the region spanning positions 5621-82574 in SEQ ID NO: 3; the region spanning positions 16120-55750 in SEQ ID NO: 4 or the region spanning positions 12473-96589 in SEQ ID NO: 5.

Detailed Description

[0023] It has been discovered that polymorphic variations in the *GP6, LAMA4, CHGB, LOC338749* and *TTN* regions described herein are associated with an increased risk of breast cancer.

[0024] The gene *GP6* (glycoprotein VI (platelet)) is also known as GPIV and GPVI. *GP6* has been mapped to chromosomal position 19q13.4. Glycoprotein VI (GP6) is a 58-kD platelet membrane glycoprotein that plays a crucial role in the collagen-induced activation and aggregation of platelets. Upon injury to the vessel wall and subsequent damage to the endothelial lining, exposure of the subendothelial matrix to blood flow results in deposition of platelets. Collagen fibers are the most thrombogenic macromolecular components of the extracellular matrix, with collagen types I, III, and VI being the major forms found in blood vessels. Platelet interaction with collagen occurs as a 2-step procedure: (1) the initial adhesion to collagen is followed by (2) an activation step leading to platelet secretion, recruitment of additional platelets, and aggregation. In physiologic conditions, the resulting platelet plug is the initial hemostatic event limiting blood loss. However, exposure of collagen after rupture of atherosclerotic plaques is a major stimulus of thrombus formation associated with myocardial

infarction or stroke. Based on the fact that GP VI is coupled to the Fc receptor-gamma chain (FCER1G; 147139) and thus should share homology with the FcR chains, they have been identified as human and mouse GP VI genes. They belong to the immunoglobulin superfamily and share 64% amino acid sequence homology. Functional evidence demonstrating the identity of the recombinant protein with GP VI was provided by binding to its natural ligand collagen; binding to convulxin (a GP VI-specific ligand from snake venom); binding of anti-GP VI IgG isolated from a patient; and association to the FcR-gamma chain.

[0025] The gene *LAMA4* (laminin, alpha 4) is also known as laminin, alpha 4 precursor and LAMA3. *LAMA4* has been mapped to chromosomal position 6q21. Laminins, a family of extracellular matrix glycoproteins, are the major noncollagenous constituent of basement membranes. They have been implicated in a wide variety of biological processes including cell adhesion, differentiation, migration, signaling, neurite outgrowth and metastasis. Laminins are composed of 3 non identical chains: laminin alpha, beta and gamma (formerly A, B1, and B2, respectively) and they form a cruciform structure consisting of 3 short arms, each formed by a different chain, and a long arm composed of all 3 chains. Each laminin chain is a multidomain protein encoded by a distinct gene. Several isoforms of each chain have been described. Different alpha, beta and gamma chain isomers combine to give rise to different heterotrimeric laminin isoforms which are designated by Arabic numerals in the order of their discovery, i.e. alpha1beta1gamma1 heterotrimer is laminin 1. The biological functions of the different chains and trimer molecules are largely unknown, but some of the chains have been shown to differ with respect to their tissue distribution, presumably reflecting diverse functions in vivo. This gene encodes the alpha chain isoform laminin, alpha 4. The domain structure of alpha 4 is similar to that of alpha 3, both of which resemble truncated versions of alpha 1 and alpha 2, in that approximately 1,200 residues at the N-terminus (domains IV, V and VI) have been lost. Laminin, alpha 4 contains the C-terminal G domain which distinguishes all alpha chains from the beta and gamma chains. The RNA analysis from adult and fetal tissues revealed developmental regulation of expression, however, the exact function of laminin, alpha 4 is not known. Tissue-specific utilization of alternative polyA-signal has been described in literature. Also, alternative splicing involving the first intron in the 5' UTR, and laminin alpha 4 like isoforms have been noted, however, the full-length nature of these products is not known.

[0026] The gene (chromogranin B (secretogranin 1) is also known as SCG1. *CHGB* has been mapped to chromosomal position 20pter-p12. Chromogranin B is a tyrosine-sulfated secretory protein found in a wide variety of peptidergic endocrine cells. Chromogranin functions as a neuroendocrine secretory granule protein which likely is the precursor for other biologically active peptides.

[0027] The gene *V20orf124* is also known as MCM8 minichromosome maintenance deficient 8 (S. cerevisiae), MGC4816, MGC12866, dJ967N21.5 and DNA replication licensing factor MCM8.

C20orf124 has been mapped to chromosomal position 20p12.3. The protein encoded by this gene is one of the highly conserved mini-chromosome maintenance proteins (MCM) that are essential for the initiation of eukaryotic genome replication. The hexameric protein complex formed by the MCM proteins is a key component of the pre-replication complex (pre_RC) and may be involved in the formation of replication forks and in the recruitment of other DNA replication related proteins. This protein contains the central domain that is conserved among the MCM proteins. This protein has been shown to co-immunoprecipitate with MCM4, 6 and 7, which suggests that it may interact with other MCM proteins and play a role in DNA replication. Alternatively spliced transcript variants encoding distinct isoforms have been described.

[0028] The gene *LOC338749* has been mapped to chromosomal position 11p15.3.

[0029] The gene *TTN* (titin) is also known as TMD, CMD1G, CMPD4, FLJ32040, connectin, CMH9, included cardiomyopathy, dilated 1G (autosomal dominant). *TTN* has been mapped to chromosomal position 2q31. This gene encodes a large abundant protein of striated muscle. The product of this gene is divided into two regions, a N-terminal I-band and a C-terminal A-band. The I-band, which is the elastic part of the molecule, contains two regions of tandem immunoglobulin domains on either side of a PEVK region that is rich in proline, glutamate, valine and lysine. The A-band, which is thought to act as a protein-ruler, contains a mixture of immunoglobulin and fibronectin repeats, and possesses kinase activity. A N-terminal Z-disc region and a C-terminal M-line region bind to the Z-line and M-line of the sarcomere respectively so that a single titin molecule spans half the length of a sarcomere. Titin also contains binding sites for muscle associated proteins so it serves as an adhesion template for the assembly of contractile machinery in muscle cells. It has also been identified as a structural protein for chromosomes. Considerable variability exists in the I-band, the M-line and the Z-disc regions of titin. Variability in the I-band region contributes to the differences in elasticity of different titin isoforms and, therefore, to the differences in elasticity of different muscle types. Of the many titin variants identified, complete transcript information is available for five. Mutations in this gene are associated with familial hypertrophic cardiomyopathy 9 and autoantibodies to titin are produced in patients with the autoimmune disease scleroderma.

Breast Cancer and Sample Selection

[0030] Breast cancer is typically described as the uncontrolled growth of malignant breast tissue. Breast cancers arise most commonly in the lining of the milk ducts of the breast (ductal carcinoma), or in the lobules where breast milk is produced (lobular carcinoma). Other forms of breast cancer include Inflammatory Breast Cancer and Recurrent Breast Cancer. Inflammatory breast cancer is a rare, but very serious, aggressive type of breast cancer. The breast may look red and feel warm with ridges, welts, or

hives on the breast; or the skin may look wrinkled. It is sometimes misdiagnosed as a simple infection. Recurrent disease means that the cancer has come back after it has been treated. It may come back in the breast, in the soft tissues of the chest (the chest wall), or in another part of the body.

[0031] As used herein, the term "breast cancer" refers to a condition characterized by anomalous rapid proliferation of abnormal cells in one or both breasts of a subject. The abnormal cells often are referred to as "neoplastic cells," which are transformed cells that can form a solid tumor. The term "tumor" refers to an abnormal mass or population of cells (*i.e.* two or more cells) that result from excessive or abnormal cell division, whether malignant or benign, and pre-cancerous and cancerous cells. Malignant tumors are distinguished from benign growths or tumors in that, in addition to uncontrolled cellular proliferation, they can invade surrounding tissues and can metastasize. In breast cancer, neoplastic cells may be identified in one or both breasts only and not in another tissue or organ, in one or both breasts and one or more adjacent tissues or organs (*e.g.* lymph node), or in a breast and one or more non-adjacent tissues or organs to which the breast cancer cells have metastasized.

[0032] The term "invasion" as used herein refers to the spread of cancerous cells to adjacent surrounding tissues. The term "invasion" often is used synonymously with the term "metastasis," which as used herein refers to a process in which cancer cells travel from one organ or tissue to another non-adjacent organ or tissue. Cancer cells in the breast(s) can spread to tissues and organs of a subject, and conversely, cancer cells from other organs or tissue can invade or metastasize to a breast. Cancerous cells from the breast(s) may invade or metastasize to any other organ or tissue of the body. Breast cancer cells often invade lymph node cells and/or metastasize to the liver, brain and/or bone and spread cancer in these tissues and organs. Breast cancers can spread to other organs and tissues and cause lung cancer, prostate cancer, colon cancer, ovarian cancer, cervical cancer, gastrointestinal cancer, pancreatic cancer, glioblastoma, bladder cancer, hepatoma, colorectal cancer, uterine cervical cancer, endometrial carcinoma, salivary gland carcinoma, kidney cancer, vulval cancer, thyroid cancer, hepatic carcinoma, skin cancer, melanoma, ovarian cancer, neuroblastoma, myeloma, various types of head and neck cancer, acute lymphoblastic leukemia, acute myeloid leukemia, Ewing sarcoma and peripheral neuroepithelioma, and other carcinomas, lymphomas, blastomas, sarcomas, and leukemias.

[0033] Breast cancers arise most commonly in the lining of the milk ducts of the breast (ductal carcinoma), or in the lobules where breast milk is produced (lobular carcinoma). Other forms of breast cancer include Inflammatory Breast Cancer and Recurrent Breast Cancer. Inflammatory Breast Cancer is a rare, but very serious, aggressive type of breast cancer. The breast may look red and feel warm with ridges, welts, or hives on the breast; or the skin may look wrinkled. It is sometimes misdiagnosed as a simple infection. Recurrent disease means that the cancer has come back after it has been treated. It may come back in the breast, in the soft tissues of the chest (the chest wall), or in another part of the body. As

used herein, the term "breast cancer" may include both Inflammatory Breast Cancer and Recurrent Breast Cancer.

[0034] In an effort to detect breast cancer as early as possible, regular physical exams and screening mammograms often are prescribed and conducted. A diagnostic mammogram often is performed to evaluate a breast complaint or abnormality detected by physical exam or routine screening mammography. If an abnormality seen with diagnostic mammography is suspicious, additional breast imaging (with exams such as ultrasound) or a biopsy may be ordered. A biopsy followed by pathological (microscopic) analysis is a definitive way to determine whether a subject has breast cancer. Excised breast cancer samples often are subjected to the following analyses: diagnosis of the breast tumor and confirmation of its malignancy; maximum tumor thickness; assessment of completeness of excision of invasive and *in situ* components and microscopic measurements of the shortest extent of clearance; level of invasion; presence and extent of regression; presence and extent of ulceration; histological type and special variants; pre-existing lesion; mitotic rate; vascular invasion; neurotropism; cell type; tumor lymphocyte infiltration; and growth phase.

[0035] The stage of a breast cancer can be classified as a range of stages from Stage 0 to Stage IV based on its size and the extent to which it has spread. The following table summarizes the stages:

Table A

Stage	Tumor Size	Lymph Node Involvement	Metastasis (Spread)
I	Less than 2 cm	No	No
II	Between 2-5 cm	No or in same side of breast	No
III	More than 5 cm	Yes, on same side of breast	No
IV	Not applicable	Not applicable	Yes

[0036] Stage 0 cancer is a contained cancer that has not spread beyond the breast ductal system. Fifteen to twenty percent of breast cancers detected by clinical examinations or testing are in Stage 0 (the earliest form of breast cancer). Two types of Stage 0 cancer are lobular carcinoma in situ (LCIS) and ductal carcinoma in situ (DCIS). LCIS indicates high risk for breast cancer. Many physicians do not classify LCIS as a malignancy and often encounter LCIS by chance on breast biopsy while investigating another area of concern. While the microscopic features of LCIS are abnormal and are similar to malignancy, LCIS does not behave as a cancer (and therefore is not treated as a cancer). LCIS is merely

a marker for a significantly increased risk of cancer anywhere in the breast. However, bilateral simple mastectomy may be occasionally performed if LCIS patients have a strong family history of breast cancer. In DCIS the cancer cells are confined to milk ducts in the breast and have not spread into the fatty breast tissue or to any other part of the body (such as the lymph nodes). DCIS may be detected on mammogram as tiny specks of calcium (known as microcalcifications) 80% of the time. Less commonly DCIS can present itself as a mass with calcifications (15% of the time); and even less likely as a mass without calcifications (<5% of the time). A breast biopsy is used to confirm DCIS. A standard DCIS treatment is breast-conserving therapy (BCT), which is lumpectomy followed by radiation treatment or mastectomy. To date, DCIS patients have chosen equally among lumpectomy and mastectomy as their treatment option, though specific cases may sometimes favor lumpectomy over mastectomy or vice versa.

[0037] In Stage I, the primary (original) cancer is 2 cm or less in diameter and has not spread to the lymph nodes. In Stage IIA, the primary tumor is between 2 and 5 cm in diameter and has not spread to the lymph nodes. In Stage IIB, the primary tumor is between 2 and 5 cm in diameter and has spread to the axillary (underarm) lymph nodes; or the primary tumor is over 5 cm and has not spread to the lymph nodes. In Stage IIIA, the primary breast cancer of any kind that has spread to the axillary (underarm) lymph nodes and to axillary tissues. In Stage IIIB, the primary breast cancer is any size, has attached itself to the chest wall, and has spread to the pectoral (chest) lymph nodes. In Stage IV, the primary cancer has spread out of the breast to other parts of the body (such as bone, lung, liver, brain). The treatment of Stage IV breast cancer focuses on extending survival time and relieving symptoms.

[0038] Based in part upon selection criteria set forth above, individuals having breast cancer can be selected for genetic studies. Also, individuals having no history of cancer or breast cancer often are selected for genetic studies. Other selection criteria can include: a tissue or fluid sample is derived from an individual characterized as Caucasian; the sample was derived from an individual of German paternal and maternal descent; the database included relevant phenotype information for the individual; case samples were derived from individuals diagnosed with breast cancer; control samples were derived from individuals free of cancer and no family history of breast cancer; and sufficient genomic DNA was extracted from each blood sample for all allelotyping and genotyping reactions performed during the study. Phenotype information included pre- or post-menopausal, familial predisposition, country or origin of mother and father, diagnosis with breast cancer (date of primary diagnosis, age of individual as of primary diagnosis, grade or stage of development, occurrence of metastases, *e.g.*, lymph node metastases, organ metastases), condition of body tissue (skin tissue, breast tissue, ovary tissue, peritoneum tissue and myometrium), method of treatment (surgery, chemotherapy, hormone therapy, radiation therapy).

[0039] Provided herein is a set of blood samples and a set of corresponding nucleic acid samples isolated from the blood samples, where the blood samples are donated from individuals diagnosed with breast cancer. The sample set often includes blood samples or nucleic acid samples from 100 or more, 150 or more, or 200 or more individuals having breast cancer, and sometimes from 250 or more, 300 or more, 400 or more, or 500 or more individuals. The individuals can have parents from any place of origin, and in an embodiment, the set of samples are extracted from individuals of German paternal and German maternal ancestry. The samples in each set may be selected based upon five or more criteria and/or phenotypes set forth above.

Polymorphic Variants Associated with Breast Cancer

[0040] A genetic analysis provided herein linked breast cancer with polymorphic variants in the *GP6, LAMA4, CHGB, LOC338749* and *TTN* regions of the human genome disclosed herein. As used herein, the term "polymorphic site" refers to a region in a nucleic acid at which two or more alternative nucleotide sequences are observed in a significant number of nucleic acid samples from a population of individuals. A polymorphic site may be a nucleotide sequence of two or more nucleotides, an inserted nucleotide or nucleotide sequence, a deleted nucleotide or nucleotide sequence, or a microsatellite, for example. A polymorphic site that is two or more nucleotides in length may be 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 or more, 20 or more, 30 or more, 50 or more, 75 or more, 100 or more, 500 or more, or about 1000 nucleotides in length, where all or some of the nucleotide sequences differ within the region. A polymorphic site is often one nucleotide in length, which is referred to herein as a "single nucleotide polymorphism" or a "SNP."

[0041] Where there are two, three, or four alternative nucleotide sequences at a polymorphic site, each nucleotide sequence is referred to as a "polymorphic variant" or "nucleic acid variant." Where two polymorphic variants exist, for example, the polymorphic variant represented in a minority of samples from a population is sometimes referred to as a "minor allele" and the polymorphic variant that is more prevalently represented is sometimes referred to as a "major allele." Many organisms possess a copy of each chromosome (*e.g.*, humans), and those individuals who possess two major alleles or two minor alleles are often referred to as being "homozygous" with respect to the polymorphism, and those individuals who possess one major allele and one minor allele are normally referred to as being "heterozygous" with respect to the polymorphism. Individuals who are homozygous with respect to one allele are sometimes predisposed to a different phenotype as compared to individuals who are heterozygous or homozygous with respect to another allele.

[0042] Furthermore, a genotype or polymorphic variant may be expressed in terms of a "haplotype," which as used herein refers to two or more polymorphic variants occurring within genomic DNA in a

group of individuals within a population. For example, two SNPs may exist within a gene where each SNP position includes a cytosine variation and an adenine variation. Certain individuals in a population may carry one allele (heterozygous) or two alleles (homozygous) having the gene with a cytosine at each SNP position. As the two cytosines corresponding to each SNP in the gene travel together on one or both alleles in these individuals, the individuals can be characterized as having a cytosine/cytosine haplotype with respect to the two SNPs in the gene.

[0043] As used herein, the term "phenotype" refers to a trait which can be compared between individuals, such as presence or absence of a condition, a visually observable difference in appearance between individuals, metabolic variations, physiological variations, variations in the function of biological molecules, and the like. An example of a phenotype is occurrence of breast cancer.

[0044] Researchers sometimes report a polymorphic variant in a database without determining whether the variant is represented in a significant fraction of a population. Because a subset of these reported polymorphic variants are not represented in a statistically significant portion of the population, some of them are sequencing errors and/or not biologically relevant. Thus, it is often not known whether a reported polymorphic variant is statistically significant or biologically relevant until the presence of the variant is detected in a population of individuals and the frequency of the variant is determined. Methods for detecting a polymorphic variant in a population are described herein, specifically in Example 2. A polymorphic variant is statistically significant and often biologically relevant if it is represented in 5% or more of a population, sometimes 10% or more, 15% or more, or 20% or more of a population, and often 25% or more, 30% or more, 35% or more, 40% or more, 45% or more, or 50% or more of a population.

[0045] A polymorphic variant may be detected on either or both strands of a double-stranded nucleic acid. For example, a thymine at a particular position in SEQ ID NO: 1 can be reported as an adenine from the complementary strand. Also, a polymorphic variant may be located within an intron or exon of a gene or within a portion of a regulatory region such as a promoter, a 5′ untranslated region (UTR), a 3′ UTR, and in DNA (*e.g.*, genomic DNA (gDNA) and complementary DNA (cDNA)), RNA (*e.g.*, mRNA, tRNA, and rRNA), or a polypeptide. Polymorphic variations may or may not result in detectable differences in gene expression, polypeptide structure, or polypeptide function.

[0046] In the genetic analysis that associated breast cancer with the polymorphic variants described hereafter, samples from individuals having breast cancer and individuals not having cancer were allelotyped and genotyped. The term "genotyped" as used herein refers to a process for determining a genotype of one or more individuals, where a "genotype" is a representation of one or more polymorphic variants in a population. Genotypes may be expressed in terms of a "haplotype," which as used herein refers to two or more polymorphic variants occurring within genomic DNA in a group of individuals within a population. For example, two SNPs may exist within a gene where each SNP position includes

a cytosine variation and an adenine variation. Certain individuals in a population may carry one allele (heterozygous) or two alleles (homozygous) having the gene with a cytosine at each SNP position. As the two cytosines corresponding to each SNP in the gene travel together on one or both alleles in these individuals, the individuals can be characterized as having a cytosine/cytosine haplotype with respect to the two SNPs in the gene.

[0047] It was determined that polymorphic variations associated with an increased risk of breast cancer existed in *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequences. Polymorphic variants in and around the *GP6, LAMA4, CHGB, LOC338749* and *TTN* loci were tested for association with breast cancer. In the *GP6* locus, these included polymorphic variants at positions in SEQ ID NO: 1 selected from the group consisting of 185, 237, 641, 719, 990, 2908, 3140, 3880, 4494, 5107, 5220, 6031, 8670, 13794, 16356, 17164, 17264, 20537, 20637, 20900, 21155, 21795, 21931, 22167, 22656, 23108, 23404, 24287, 24480, 24592, 24878, 26370, 27056, 27874, 31248, 31458, 31553, 31637, 31668, 31752, 37643, 43941, 44134, 44329, 44343, 44362, 44818, 44917, 45215, 45666, 45680, 46402 ,46510, 46554 ,46823, 47714, 48963, 49157, 49254, 49257, 49356, 55202, 55527, 55916, 56402, 56413, 56685, 56783, 58044, ,58301 ,58382, 58393, 58869, 59155, 59189, 62546, 62568, 70983, 71465, 71538, 72144, 72340, 72527, 72968, 73397, 73553, 73720, 74190, 74687, 74699, 75580, 76345, 76506, 77554, 77889, 77919, 78866, 79061, 83777, 84360, 84631, 85775, 87153, 89650, 89895, 90103, 90234, 90309, 90376, 90925, 91561, 91605, 92954 and 94228. Polymorphic variants in a region spanning positions 185-8377 in SEQ ID NO: 1 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 185, 20537, 44329, 44362, 45666, 45680, 46510, 49254, 49356, 56402, 58301, 71465, 72527, 73553, 76345 and 83777 in SEQ ID NO: 1. At these positions in SEQ ID NO: 1, a thymine at position 185, a thymine at position 20537, a cytosine at position 44329, an adenine at position 44362, a guanine at position 45666, a thymine at position 45680, a thymine at position 46510, a guanine at position 49254, a thymine at position 49356, a guanine at position 56402, a cytosine at position 58301, an adenine at position 71465, a guanine at position 72527, a guanine at position 73553, a thymine at position 76345 and an adenine at position 83777 in particular were associated with risk of breast cancer. Also, a lysine at amino acid position 237 in SEQ ID NO: 12 was associated with an increased risk of breast cancer.

[0048] In the *LAMA4* locus, these included polymorphic variants at positions in SEQ ID NO: 2 selected from the group consisting of 184, 506, 3981, 7815, 7875, 10775, 10786, 11013, 11020, 11101, 14171, 14278, 16512, 16706, 18442, 20286, 21591, 22275, 25318, 27997, 29840,31088, 31258, 32367, 32427, 33671, 38796, 41530, 41874, 44161, 47502, 51089, 51205, 53645, 54280, 57610, 57740, 60812, 60837, 64448, 65249, 65482, 66535, 66789, 67214, 68347, 69060, 70100, 70215, 73687, 73732, 74183, 74813, 78136, 79540, 79655, 79731, 82111, 82155, 83479, 84511, 85290, 90620, 91127, 92095, 92679,

94839 and 95220. Polymorphic variants in a region spanning positions 506-95220 in SEQ ID NO: 2 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 506, 3981, 7815, 7875, 11020, 11101, 18442, 47502, 53645, 65249, 73687, 73732, 74183, 79540, 82155, 85290, 90620, 91127, 92095, 92679, 94839 and 95220 in SEQ ID NO: 2. At these positions in SEQ ID NO: 2, a cytosine at position 506, a cytosine at position 3981, a guanine at position 7815, a guanine at position 7875, a thymine at position 11020, an adenine at position 11101, an adenine at position 18442, a cytosine at position 47502, a guanine at position 53645, a thymine at position 65249, a cytosine at position 73687, an adenine at position 73732, a thymine at position 74183, a thymine at position 79540, a thymine at position 82155, a cytosine at position 85290, a guanine at position 90620, a guanine at position 91127, an adenine at position 92095, a guanine at position 92679, a guanine at position 94839 and a cytosine at position 95220 in particular were associated with increased risk of breast cancer.

[0049] In the *CHGB* locus, these included polymorphic variants at positions in SEQ ID NO: 3 selected from the group consisting of 186, 1332, 1893, 2786, 2962, 3377, 5522, 5621, 5889, 7531, 8268, 8923, 8988, 9117, 9448, 9494, 9628, 9640, 11072, 11150, 11379, 11692, 12056, 12104, 14160, 14836, 14980, 15165, 15315, 15624, 15796, 15939, 16581, 17045, 18501, 21800, 21966, 22134, 22181, 23028, 23312, 23573, 23858, 23888, 23990, 24073, 25330, 26473, 27958, 28421, 28804, 29322, 30819, 31956, 32592, 32818, 32880, 33244, 33845, 34272, 34931, 36870, 37790, 38708, 39135, 39919, 40166, 40985, 41049, 41935, 42775, 43807, 44254, 44814, 45249, 47599, 47807, 48555, 49249, 49293, 57566, 63587, 64560, 65432, 66291, 71331, 73344, 74159, 74564, 78194, 79128, 79393, 81579, 82574, 85309, 87076, 87844 and 90241. Polymorphic variants in a region spanning positions 5621-82574 in SEQ ID NO: 3 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 5621, 9628, 9640, 21800, 21966, 22134, 22181, 23028, 23573, 23888, 24073, 26473, 27958, 28421, 28804, 29322, 30819, 31956, 32592, 32818, 32880, 33244, 33845, 34931, 37790, 38708, 39135, 39919, 40166, 41049, 43807, 44254, 45249, 47807, 48555, 49249, 49293, 57566, 63587, 64560, 65432, 66291, 71331, 73344, 74159, 78194, 79128, 81579 and 82574 in SEQ ID NO: 3. At these positions in SEQ ID NO: 3, a guanine at position 5621, a guanine at position 9628, a cytosine at position 9640, a guanine at position 21800, an adenine at position 21966, a guanine at position 22134, an adenine at position 22181, a guanine at position 23028, a thymine at position 23573, a guanine at positino 23888, an adenine at position 24073, a thymine at position 26473, a cytosine at position 27958, an adenine at position 28421, a thymine at position 28804, a cytosine at position 29322, a cytosine at position 30819, a guanine at position 31956, a guanine at position 32592, a cytosine at position 32818, a thymine at position 32880, a cytosine at position 33244, an adenine at position 33845, a thymine at position 34931, a thymine at position 37790, a guanine at position 38708, a thymine at position 39135, an adenine at

position 39919, a thymine at position 40166, a guanine at position 41049, a cytosine at position 43807, a guanine at position 44254, a thymine at position 45249, a guanine at position 47807, a cytosine at position 48555, an adenine at position 49249, a cytosine at position 49293, a cytosine at position 57566, a cytosine at position 63587, a thymine at position 64560, a cytosine at position 65432, a thymine at position 66291, an adenine at position 71331, a thymine at position 73344, a thymine at position 74159, an adenine at position 78194, a cytosine at position 79128, an adenine at position 81579 and a cytosine at position 82574 in particular were associated with an increased risk of breast cancer. Also, a proline at amino acid position 413 and a glutamine at amino acid position 63 in particular associated with an increased risk of breast cancer.

[0050] In the *LOC338749* locus, these included polymorphic variants at positions in SEQ ID NO: 4 selected from the group consisting of 142, 693, 731, 879, 1084, 2249, 2519, 4461, 4616, 5109, 5270, 5436, 5457, 6536, 9665, 16120, 29489, 29524, 49159, 49273, 49596, 50135, 50184, 50393, 50401, 55750, 73843, 73852, 74052, 75382, 75662, 75942, 77917, 78821, 94813 and 97149. Polymorphic variants in a region spanning positions 16120-55750 in SEQ ID NO: 4 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 16120 and 55750 in SEQ ID NO: 4. At these positions in SEQ ID NO: 4, a thymine at position 16120 and a cytosine at position 55750 in particular were associated with an increased risk of breast cancer.

[0051] In the *TTN* locus, these included polymorphic variants at positions in SEQ ID NO: 5 selected from the group consisting of 200, 381, 5303, 6084, 6879, 7837, 7985, 9333, 11559, 12473, 12880, 13606, 14861, 20658, 22200, 24525, 26373, 42869, 43713, 44429, 49037, 49170, 50206, 51552, 51674, 56427, 56844, 57953, 60862, 61606, 62560, 65078, 65155, 70295, 70335, 70398, 79233, 80025, 84521, 84540, 85170, 85300, 87596, 89696, 92219 and 96589. Polymorphic variants in a region spanning positions 12473-96589 in SEQ ID NO: 5 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 12473, 20658, 24525, 49037, 49170, 51552, 51674, 70335, 84521, 87596, 92219 and 96589 in SEQ ID NO: 5. At these positions in SEQ ID NO: 5, a cytosine at position 12473, a thymine at position 20658, a cytosine at position 24525, a guanine at position 49037, an adenine at position 49170, a thymine at position 51552, a guanine at position 51674, a thymine at position 70335, a guanine at position 84521, an adenine at position 87596, an adenine at position 92219 and a thymine at position 96589 in particular were associated with an increased risk of breast cancer.

Additional Polymorphic Variants Associated with Breast Cancer

[0052] Also provided is a method for identifying polymorphic variants proximal to an incident, founder polymorphic variant associated with breast cancer. Thus, featured herein are methods for

identifying a polymorphic variation associated with breast cancer that is proximal to an incident polymorphic variation associated with breast cancer, which comprises identifying a polymorphic variant proximal to the incident polymorphic variant associated with breast cancer, where the incident polymorphic variant is in a nucleotide sequence set forth in SEQ ID NO: 1-5. The nucleotide sequence often comprises a polynucleotide sequence selected from the group consisting of (a) a nucleotide sequence set forth in SEQ ID NO: 1-5; (b) a nucleotide sequence which encodes a polypeptide having an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5; (c) a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5 or a nucleotide sequence about 90% or more identical to the nucleotide sequence set forth in SEQ ID NO: 1-5; and (d) a fragment of a nucleotide sequence of (a), (b), or (c), often a fragment that includes a polymorphic site associated with breast cancer. The presence or absence of an association of the proximal polymorphic variant with breast cancer then is determined using a known association method, such as a method described in the Examples hereafter. In an embodiment, the incident polymorphic variant is described in SEQ ID NO: 1-5. In another embodiment, the proximal polymorphic variant identified sometimes is a publicly disclosed polymorphic variant, which for example, sometimes is published in a publicly available database. In other embodiments, the polymorphic variant identified is not publicly disclosed and is discovered using a known method, including, but not limited to, sequencing a region surrounding the incident polymorphic variant in a group of nucleic acid samples. Thus, multiple polymorphic variants proximal to an incident polymorphic variant are associated with breast cancer using this method.

[0053] The proximal polymorphic variant often is identified in a region surrounding the incident polymorphic variant. In certain embodiments, this surrounding region is about 50 kb flanking the first polymorphic variant (*e.g.* about 50 kb 5' of the first polymorphic variant and about 50 kb 3' of the first polymorphic variant), and the region sometimes is composed of shorter flanking sequences, such as flanking sequences of about 40 kb, about 30 kb, about 25 kb, about 20 kb, about 15 kb, about 10 kb, about 7 kb, about 5 kb, or about 2 kb 5' and 3' of the incident polymorphic variant. In other embodiments, the region is composed of longer flanking sequences, such as flanking sequences of about 55 kb, about 60 kb, about 65 kb, about 70 kb, about 75 kb, about 80 kb, about 85 kb, about 90 kb, about 95 kb, or about 100 kb 5' and 3' of the incident polymorphic variant.

[0054] In certain embodiments, polymorphic variants associated with breast cancer are identified iteratively. For example, a first proximal polymorphic variant is associated with breast cancer using the methods described above and then another polymorphic variant proximal to the first proximal polymorphic variant is identified (*e.g.*, publicly disclosed or discovered) and the presence or absence of

an association of one or more other polymorphic variants proximal to the first proximal polymorphic variant with breast cancer is determined.

[0055] The methods described herein are useful for identifying or discovering additional polymorphic variants that may be used to further characterize a gene, region or loci associated with a condition, a disease (*e.g.*, breast cancer), or a disorder. For example, allelotyping or genotyping data from the additional polymorphic variants may be used to identify a functional mutation or a region of linkage disequilibrium.

[0056] In certain embodiments, polymorphic variants identified or discovered within a region comprising the first polymorphic variant associated with breast cancer are genotyped using the genetic methods and sample selection techniques described herein, and it can be determined whether those polymorphic variants are in linkage disequilibrium with the first polymorphic variant. The size of the region in linkage disequilibrium with the first polymorphic variant also can be assessed using these genotyping methods. Thus, provided herein are methods for determining whether a polymorphic variant is in linkage disequilibrium with a first polymorphic variant associated with breast cancer, and such information can be used in prognosis methods described herein.

Isolated *GP6, LAMA4, CHGB, LOC338749* or *TTN* Nucleic Acids

[0057] Featured herein are isolated *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acids, which include the nucleic acid having the nucleotide sequence of SEQ ID NO: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or 11, nucleic acid variants, and substantially identical nucleic acids of the foregoing. Nucleotide sequences of the *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acids sometimes are referred to herein as "*GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequences." A "*GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid variant" refers to one allele that may have one or more different polymorphic variations as compared to another allele in another subject or the same subject. A polymorphic variation in the *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid variant may be represented on one or both strands in a double-stranded nucleic acid or on one chromosomal complement (heterozygous) or both chromosomal complements (homozygous).

[0058] As used herein, the term "nucleic acid" includes DNA molecules (e.g., a complementary DNA (cDNA) and genomic DNA (gDNA)) and RNA molecules (e.g., mRNA, rRNA, and tRNA) and analogs of DNA or RNA, for example, by use of nucleotide analogs. The nucleic acid molecule can be single-stranded and it is often double-stranded. The term "isolated or purified nucleic acid" refers to nucleic acids that are separated from other nucleic acids present in the natural source of the nucleic acid. For example, with regard to genomic DNA, the term "isolated" includes nucleic acids which are separated from the chromosome with which the genomic DNA is naturally associated. An "isolated"

nucleic acid is often free of sequences which naturally flank the nucleic acid (i.e., sequences located at the 5' and/or 3' ends of the nucleic acid) in the genomic DNA of the organism from which the nucleic acid is derived. For example, in various embodiments, the isolated nucleic acid molecule can contain less than about 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb or 0.1 kb of 5' and/or 3' nucleotide sequences which flank the nucleic acid molecule in genomic DNA of the cell from which the nucleic acid is derived. Moreover, an "isolated" nucleic acid molecule, such as a cDNA molecule, can be substantially free of other cellular material, or culture medium when produced by recombinant techniques, or substantially free of chemical precursors or other chemicals when chemically synthesized. As used herein, the term "*GP6, LAMA4, CHGB, LOC338749* or *TTN* gene" refers to a nucleotide sequence that encodes a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide.

[0059] Also included herein are nucleic acid fragments. These fragments typically are a nucleotide sequence identical to a nucleotide sequence in SEQ ID NO: 1-11, a nucleotide sequence substantially identical to a nucleotide sequence in SEQ ID NO: 1-11, or a nucleotide sequence that is complementary to the foregoing. The nucleic acid fragment may be identical, substantially identical or homologous to a nucleotide sequence in an exon or an intron in SEQ ID NO: 1-5, and may encode a domain or part of a domain or motif of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide, sometimes the domains set forth in Figures 13-18. Sometimes, the fragment comprises the polymorphic variation described herein as being associated with breast cancer. The nucleic acid fragment sometimes is 50, 100, or 200 or fewer base pairs in length, and is sometimes about 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3800, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 15000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 110000, 120000, 130000, 140000, 150000 or 160000 base pairs in length. A nucleic acid fragment complementary to a nucleotide sequence identical or substantially identical to the nucleotide sequence of SEQ ID NO: 1-11 and hybridizes to such a nucleotide sequence under stringent conditions often is referred to as a "probe." Nucleic acid fragments often include one or more polymorphic sites, or sometimes have an end that is adjacent to a polymorphic site as described hereafter.

[0060] An example of a nucleic acid fragment is an oligonucleotide. As used herein, the term "oligonucleotide" refers to a nucleic acid comprising about 8 to about 50 covalently linked nucleotides, often comprising from about 8 to about 35 nucleotides, and more often from about 10 to about 25 nucleotides. The backbone and nucleotides within an oligonucleotide may be the same as those of naturally occurring nucleic acids, or analogs or derivatives of naturally occurring nucleic acids, provided that oligonucleotides having such analogs or derivatives retain the ability to hybridize specifically to a nucleic acid comprising a targeted polymorphism. Oligonucleotides described herein may be used as

hybridization probes or as components of prognostic or diagnostic assays, for example, as described herein.

[0061] Oligonucleotides are typically synthesized using standard methods and equipment, such as the ABI 3900 High Throughput DNA Synthesizer and the EXPEDITE™ 8909 Nucleic Acid Synthesizer, both of which are available from Applied Biosystems (Foster City, CA). Analogs and derivatives are exemplified in U.S. Pat. Nos. 4,469,863; 5,536,821; 5,541,306; 5,637,683; 5,637,684; 5,700,922; 5,717,083; 5,719,262; 5,739,308; 5,773,601; 5,886,165; 5,929,226; 5,977,296; 6,140,482; WO 00/56746; WO 01/14398, and related publications. Methods for synthesizing oligonucleotides comprising such analogs or derivatives are disclosed, for example, in the patent publications cited above and in U.S. Pat. Nos. 5,614,622; 5,739,314; 5,955,599; 5,962,674; 6,117,992; in WO 00/75372; and in related publications.

[0062] Oligonucleotides also may be linked to a second moiety. The second moiety may be an additional nucleotide sequence such as a tail sequence (e.g., a polyadenosine tail), an adapter sequence (e.g., phage M13 universal tail sequence), and others. Alternatively, the second moiety may be a non-nucleotide moiety such as a moiety which facilitates linkage to a solid support or a label to facilitate detection of the oligonucleotide. Such labels include, without limitation, a radioactive label, a fluorescent label, a chemiluminescent label, a paramagnetic label, and the like. The second moiety may be attached to any position of the oligonucleotide, provided the oligonucleotide can hybridize to the nucleic acid comprising the polymorphism.

Uses for Nucleic Acid Sequences

[0063] Nucleic acid coding sequences depicted in SEQ ID NO: 1-11 may be used for diagnostic purposes for detection and control of polypeptide expression. Also, included herein are oligonucleotide sequences such as antisense RNA, small-interfering RNA (siRNA) and DNA molecules and ribozymes that function to inhibit translation of a polypeptide. Antisense techniques and RNA interference techniques are known in the art and are described herein.

[0064] Ribozymes are enzymatic RNA molecules capable of catalyzing the specific cleavage of RNA. The mechanism of ribozyme action involves sequence specific hybridization of the ribozyme molecule to complementary target RNA, followed by a endonucleolytic cleavage. Ribozymes may be engineered hammerhead motif ribozyme molecules that specifically and efficiently catalyze endonucleolytic cleavage of RNA sequences corresponding to or complementary to the nucleotide sequences set forth in SEQ ID NO: 1-11. Specific ribozyme cleavage sites within any potential RNA target are initially identified by scanning the target molecule for ribozyme cleavage sites which include the following sequences, GUA, GUU and GUC. Once identified, short RNA sequences of between

fifteen (15) and twenty (20) ribonucleotides corresponding to the region of the target gene containing the cleavage site may be evaluated for predicted structural features such as secondary structure that may render the oligonucleotide sequence unsuitable. The suitability of candidate targets may also be evaluated by testing their accessibility to hybridization with complementary oligonucleotides, using ribonuclease protection assays.

[0065] Antisense RNA and DNA molecules, siRNA and ribozymes may be prepared by any method known in the art for the synthesis of RNA molecules. These include techniques for chemically synthesizing oligodeoxyribonucleotides well known in the art such as solid phase phosphoramidite chemical synthesis. Alternatively, RNA molecules may be generated by *in vitro* and *in vivo* transcription of DNA sequences encoding the antisense RNA molecule. Such DNA sequences may be incorporated into a wide variety of vectors which incorporate suitable RNA polymerase promoters such as the T7 or SP6 polymerase promoters. Alternatively, antisense cDNA constructs that synthesize antisense RNA constitutively or inducibly, depending on the promoter used, can be introduced stably into cell lines.

[0066] DNA encoding a polypeptide also may have a number of uses for the diagnosis of diseases, including breast cancer, resulting from aberrant expression of a target gene described herein. For example, the nucleic acid sequence may be used in hybridization assays of biopsies or autopsies to diagnose abnormalities of expression or function (*e.g.*, Southern or Northern blot analysis, in situ hybridization assays).

[0067] In addition, the expression of a polypeptide during embryonic development may also be determined using nucleic acid encoding the polypeptide. As addressed, *infra*, production of functionally impaired polypeptide can be the cause of various disease states, such as breast cancer. *In situ* hybridizations using polynucleotide probes may be employed to predict problems related to breast cancer. Further, as indicated, *infra*, administration of human active polypeptide, recombinantly produced as described herein, may be used to treat disease states related to functionally impaired polypeptide. Alternatively, gene therapy approaches may be employed to remedy deficiencies of functional polypeptide or to replace or compete with dysfunctional polypeptide.

Expression Vectors, Host Cells, and Genetically Engineered Cells

[0068] Provided herein are nucleic acid vectors, often expression vectors, which contain a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid. As used herein, the term "vector" refers to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked and can include a plasmid, cosmid, or viral vector. The vector can be capable of autonomous replication or it can integrate into a host DNA. Viral vectors may include replication defective retroviruses, adenoviruses and adeno-associated viruses for example.

[0069] A vector can include a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid in a form suitable for expression of the nucleic acid in a host cell. The recombinant expression vector typically includes one or more regulatory sequences operatively linked to the nucleic acid sequence to be expressed. The term "regulatory sequence" includes promoters, enhancers and other expression control elements (e.g., polyadenylation signals). Regulatory sequences include those that direct constitutive expression of a nucleotide sequence, as well as tissue-specific regulatory and/or inducible sequences. The design of the expression vector can depend on such factors as the choice of the host cell to be transformed, the level of expression of polypeptide desired, and the like. Expression vectors can be introduced into host cells to produce *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides, including fusion polypeptides, encoded by *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acids.

[0070] Recombinant expression vectors can be designed for expression of *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides in prokaryotic or eukaryotic cells. For example, *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides can be expressed in E. coli, insect cells (e.g., using baculovirus expression vectors), yeast cells, or mammalian cells. Suitable host cells are discussed further in Goeddel, Gene Expression Technology: Methods in Enzymology 185, Academic Press, San Diego, CA (1990). Alternatively, the recombinant expression vector can be transcribed and translated in vitro, for example using T7 promoter regulatory sequences and T7 polymerase.

[0071] Expression of polypeptides in prokaryotes is most often carried out in E. coli with vectors containing constitutive or inducible promoters directing the expression of either fusion or non-fusion polypeptides. Fusion vectors add a number of amino acids to a polypeptide encoded therein, usually to the amino terminus of the recombinant polypeptide. Such fusion vectors typically serve three purposes: 1) to increase expression of recombinant polypeptide; 2) to increase the solubility of the recombinant polypeptide; and 3) to aid in the purification of the recombinant polypeptide by acting as a ligand in affinity purification. Often, a proteolytic cleavage site is introduced at the junction of the fusion moiety and the recombinant polypeptide to enable separation of the recombinant polypeptide from the fusion moiety subsequent to purification of the fusion polypeptide. Such enzymes, and their cognate recognition sequences, include Factor Xa, thrombin and enterokinase. Typical fusion expression vectors include pGEX (Pharmacia Biotech Inc; Smith & Johnson, Gene 67: 31-40 (1988)), pMAL (New England Biolabs, Beverly, MA) and pRIT5 (Pharmacia, Piscataway, NJ) which fuse glutathione S-transferase (GST), maltose E binding polypeptide, or polypeptide A, respectively, to the target recombinant polypeptide.

[0072] Purified fusion polypeptides can be used in screening assays and to generate antibodies specific for *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides. In a therapeutic embodiment, fusion polypeptide expressed in a retroviral expression vector is used to infect bone marrow cells that are

subsequently transplanted into irradiated recipients. The pathology of the subject recipient is then examined after sufficient time has passed (e.g., six (6) weeks).

[0073] Expressing the polypeptide in host bacteria with an impaired capacity to proteolytically cleave the recombinant polypeptide is often used to maximize recombinant polypeptide expression (Gottesman, S., Gene Expression Technology: Methods in Enzymology, Academic Press, San Diego, California 185: 119-128 (1990)). Another strategy is to alter the nucleotide sequence of the nucleic acid to be inserted into an expression vector so that the individual codons for each amino acid are those preferentially utilized in E. coli (Wada et al., Nucleic Acids Res. 20: 2111-2118 (1992)). Such alteration of nucleotide sequences can be carried out by standard DNA synthesis techniques.

[0074] When used in mammalian cells, the expression vector's control functions are often provided by viral regulatory elements. For example, commonly used promoters are derived from polyoma, Adenovirus 2, cytomegalovirus and Simian Virus 40. Recombinant mammalian expression vectors are often capable of directing expression of the nucleic acid in a particular cell type (e.g., tissue-specific regulatory elements are used to express the nucleic acid). Non-limiting examples of suitable tissue-specific promoters include an albumin promoter (liver-specific; Pinkert et al., Genes Dev. 1: 268-277 (1987)), lymphoid-specific promoters (Calame & Eaton, Adv. Immunol. 43: 235-275 (1988)), promoters of T cell receptors (Winoto & Baltimore, EMBO J. 8: 729-733 (1989)) promoters of immunoglobulins (Banerji et al., Cell 33: 729-740 (1983); Queen & Baltimore, Cell 33: 741-748 (1983)), neuron-specific promoters (e.g., the neurofilament promoter; Byrne & Ruddle, Proc. Natl. Acad. Sci. USA 86: 5473-5477 (1989)), pancreas-specific promoters (Edlund et al., Science 230: 912-916 (1985)), and mammary gland-specific promoters (e.g., milk whey promoter; U.S. Patent No. 4,873,316 and European Application Publication No. 264,166). Developmentally-regulated promoters are sometimes utilized, for example, the murine hox promoters (Kessel & Gruss, Science 249: 374-379 (1990)) and the α-fetopolypeptide promoter (Campes & Tilghman, Genes Dev. 3: 537-546 (1989)).

[0075] A *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid may also be cloned into an expression vector in an antisense orientation. Regulatory sequences (e.g., viral promoters and/or enhancers) operatively linked to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid cloned in the antisense orientation can be chosen for directing constitutive, tissue specific or cell type specific expression of antisense RNA in a variety of cell types. Antisense expression vectors can be in the form of a recombinant plasmid, phagemid or attenuated virus. For a discussion of the regulation of gene expression using antisense genes see Weintraub et al., Antisense RNA as a molecular tool for genetic analysis, Reviews - Trends in Genetics, Vol. 1(1) (1986).

[0076] Also provided herein are host cells that include a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid within a recombinant expression vector or *GP6, LAMA4, CHGB, LOC338749* or *TTN*

nucleic acid sequence fragments which allow it to homologously recombine into a specific site of the host cell genome. The terms "host cell" and "recombinant host cell" are used interchangeably herein. Such terms refer not only to the particular subject cell but rather also to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein. A host cell can be any prokaryotic or eukaryotic cell. For example, a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide can be expressed in bacterial cells such as E. coli, insect cells, yeast or mammalian cells (such as Chinese hamster ovary cells (CHO) or COS cells). Other suitable host cells are known to those skilled in the art.

[0077] Vectors can be introduced into host cells via conventional transformation or transfection techniques. As used herein, the terms "transformation" and "transfection" are intended to refer to a variety of art-recognized techniques for introducing foreign nucleic acid (e.g., DNA) into a host cell, including calcium phosphate or calcium chloride co-precipitation, transduction/infection, DEAE-dextran-mediated transfection, lipofection, or electroporation.

[0078] A host cell provided herein can be used to produce (i.e., express) a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. Accordingly, further provided are methods for producing a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide using the host cells described herein. In one embodiment, the method includes culturing host cells into which a recombinant expression vector encoding a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide has been introduced in a suitable medium such that a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide is produced. In another embodiment, the method further includes isolating a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide from the medium or the host cell.

[0079] Also provided are cells or purified preparations of cells which include a *GP6, LAMA4, CHGB, LOC338749* or *TTN* transgene, or which otherwise misexpress *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. Cell preparations can consist of human or non-human cells, e.g., rodent cells, e.g., mouse or rat cells, rabbit cells, or pig cells. In certain embodiments, the cell or cells include a *GP6, LAMA4, CHGB, LOC338749* or *TTN* transgene (e.g., a heterologous form of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* such as a human gene expressed in non-human cells). The *GP6, LAMA4, CHGB, LOC338749* or *TTN* transgene can be misexpressed, e.g., overexpressed or underexpressed. In other embodiments, the cell or cells include a gene which misexpress an endogenous *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide (e.g., expression of a gene is disrupted, also known as a knockout). Such cells can serve as a model for studying disorders which are related to mutated or mis-expressed *GP6, LAMA4, CHGB, LOC338749* or *TTN* alleles or for use in drug screening. Also provided

are human cells (e.g., a hematopoietic stem cells) transformed with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid.

[0080] Also provided are cells or a purified preparation thereof (e.g., human cells) in which an endogenous *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid is under the control of a regulatory sequence that does not normally control the expression of the endogenous *GP6, LAMA4, CHGB, LOC338749* or *TTN* gene. The expression characteristics of an endogenous gene within a cell (e.g., a cell line or microorganism) can be modified by inserting a heterologous DNA regulatory element into the genome of the cell such that the inserted regulatory element is operably linked to the endogenous *GP6, LAMA4, CHGB, LOC338749* or *TTN* gene. For example, an endogenous *GP6, LAMA4, CHGB, LOC338749* or *TTN* gene (e.g., a gene which is "transcriptionally silent," not normally expressed, or expressed only at very low levels) may be activated by inserting a regulatory element which is capable of promoting the expression of a normally expressed gene product in that cell. Techniques such as targeted homologous recombinations, can be used to insert the heterologous DNA as described in, e.g., Chappel, US 5,272,071; WO 91/06667, published on May 16, 1991.

Transgenic Animals

[0081] Non-human transgenic animals that express a heterologous *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide (e.g., expressed from a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid isolated from another organism) can be generated. Such animals are useful for studying the function and/or activity of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide and for identifying and/or evaluating modulators of *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid and *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide activity. As used herein, a "transgenic animal" is a non-human animal such as a mammal (e.g., a non-human primate such as chimpanzee, baboon, or macaque; an ungulate such as an equine, bovine, or caprine; or a rodent such as a rat, a mouse, or an Israeli sand rat), a bird (e.g., a chicken or a turkey), an amphibian (e.g., a frog, salamander, or newt), or an insect (e.g., Drosophila melanogaster), in which one or more of the cells of the animal includes a *GP6, LAMA4, CHGB, LOC338749* or *TTN* transgene. A transgene is exogenous DNA or a rearrangement (e.g., a deletion of endogenous chromosomal DNA) that is often integrated into or occurs in the genome of cells in a transgenic animal. A transgene can direct expression of an encoded gene product in one or more cell types or tissues of the transgenic animal, and other transgenes can reduce expression (e.g., a knockout). Thus, a transgenic animal can be one in which an endogenous *GP6, LAMA4, CHGB, LOC338749* or *TTN* gene has been altered by homologous recombination between the endogenous gene and an exogenous DNA molecule introduced into a cell of the animal (e.g., an embryonic cell of the animal) prior to development of the animal.

[0082] Intronic sequences and polyadenylation signals can also be included in the transgene to increase expression efficiency of the transgene. One or more tissue-specific regulatory sequences can be operably linked to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* transgene to direct expression of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide to particular cells. A transgenic founder animal can be identified based upon the presence of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* transgene in its genome and/or expression of *GP6, LAMA4, CHGB, LOC338749* or *TTN* mRNA in tissues or cells of the animals. A transgenic founder animal can then be used to breed additional animals carrying the transgene. Moreover, transgenic animals carrying a transgene encoding a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide can further be bred to other transgenic animals carrying other transgenes.

[0083] *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides can be expressed in transgenic animals or plants by introducing, for example, a nucleic acid encoding the polypeptide into the genome of an animal. In certain embodiments the nucleic acid is placed under the control of a tissue specific promoter, e.g., a milk or egg specific promoter, and recovered from the milk or eggs produced by the animal. Also included is a population of cells from a transgenic animal.

GP6, LAMA4, CHGB, LOC338749 and *TTN* Polypeptides

[0084] Featured herein are isolated *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides, which include polypeptides having amino acid sequences set forth in SEQ ID NO: 12-17, and substantially identical polypeptides thereof. Such polypeptides sometimes are proteins or peptides. A *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide is a polypeptide encoded by a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid, where one nucleic acid can encode one or more different polypeptides. An "isolated" or "purified" polypeptide or protein is substantially free of cellular material or other contaminating proteins from the cell or tissue source from which the protein is derived, or substantially free from chemical precursors or other chemicals when chemically synthesized. In one embodiment, the language "substantially free" means preparation of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide variant having less than about 30%, 20%, 10% and sometimes 5% (by dry weight), of non-*GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide (also referred to herein as a "contaminating protein"), or of chemical precursors or non-*GP6, LAMA4, CHGB, LOC338749* or *TTN* chemicals. When the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or a biologically active portion thereof is recombinantly produced, it is also often substantially free of culture medium, specifically, where culture medium represents less than about 20%, sometimes less than about 10%, and often less than about 5% of the volume of the polypeptide preparation. Isolated or purified *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide preparations are

sometimes 0.01 milligrams or more or 0.1 milligrams or more, and often 1.0 milligrams or more and 10 milligrams or more in dry weight. In specific embodiments, a GP6 polypeptide comprises a lysine at amino acid position 237 of SEQ ID NO: 12, and a CHGB polypeptide comprises a proline at amino acid position 413 of SEQ ID NO: 16 or a glutamine at amino acid position 63 of SEQ ID NO: 16.

[0085] In another aspect, featured herein are *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides and biologically active or antigenic fragments thereof that are useful as reagents or targets in assays applicable to prevention, treatment or diagnosis of breast cancer. In another embodiment, provided herein are *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides having a *GP6, LAMA4, CHGB, LOC338749* or *TTN* activity or activities.

[0086] Further included herein are *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide fragments. The polypeptide fragment may be a domain or part of a domain of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. The polypeptide fragment is often 50 or fewer, 100 or fewer, or 200 or fewer amino acids in length, and is sometimes 300, 400, 500, 600, 700, or 900 or fewer amino acids in length. In certain embodiments, the polypeptide fragment comprises, consists essentially of, or consists of, at least 6 consecutive amino acids and not more than 1211 consecutive amino acids of SEQ ID NO: 12-17, or the polypeptide fragment comprises, consists essentially of, or consists of, at least 6 consecutive amino acids and not more than 543 consecutive amino acids of SEQ ID NO: 12-17.

[0087] *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides described herein can be used as immunogens to produce anti-*GP6, LAMA4, CHGB, LOC338749* or *TTN* antibodies in a subject, to purify *GP6, LAMA4, CHGB, LOC338749* or *TTN* ligands or binding partners, and in screening assays to identify molecules which inhibit or enhance the interaction of *GP6, LAMA4, CHGB, LOC338749* or *TTN* with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* substrate. Full-length *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides and polynucleotides encoding the same may be specifically substituted for a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide fragment or polynucleotide encoding the same in any embodiment described herein.

[0088] Substantially identical polypeptides may depart from the amino acid sequences set forth in SEQ ID NO: 12-17 in different manners. For example, conservative amino acid modifications may be introduced at one or more positions in the amino acid sequences of SEQ ID NO: 12-17. A "conservative amino acid substitution" is one in which the amino acid is replaced by another amino acid having a similar structure and/or chemical function. Families of amino acid residues having similar structures and functions are well known. These families include amino acids with basic side chains (e.g., lysine, arginine, histidine), acidic side chains (e.g., aspartic acid, glutamic acid), uncharged polar side chains (e.g., glycine, asparagine, glutamine, serine, threonine, tyrosine, cysteine), nonpolar side chains (e.g., alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan), beta-branched side

chains (e.g., threonine, valine, isoleucine) and aromatic side chains (e.g., tyrosine, phenylalanine, tryptophan, histidine). Also, essential and non-essential amino acids may be replaced. A "non-essential" amino acid is one that can be altered without abolishing or substantially altering the biological function of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide, whereas altering an "essential" amino acid abolishes or substantially alters the biological function of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. Amino acids that are conserved among *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides are typically essential amino acids.

[0089] Also, *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides and polypeptide variants may exist as chimeric or fusion polypeptides. As used herein, a *GP6, LAMA4, CHGB, LOC338749* or *TTN* "chimeric polypeptide" or "fusion polypeptide" includes a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide linked to a non-*GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. A "non-*GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide" refers to a polypeptide having an amino acid sequence corresponding to a polypeptide which is not substantially identical to the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide, which includes, for example, a polypeptide that is different from the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide and derived from the same or a different organism. The *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide in the fusion polypeptide can correspond to an entire or nearly entire *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or a fragment thereof. The non-*GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide can be fused to the N-terminus or C-terminus of the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide.

[0090] Fusion polypeptides can include a moiety having high affinity for a ligand. For example, the fusion polypeptide can be a GST-*GP6, LAMA4, CHGB, LOC338749* or *TTN* fusion polypeptide in which the *GP6, LAMA4, CHGB, LOC338749* or *TTN* sequences are fused to the C-terminus of the GST sequences, or a polyhistidine-*GP6, LAMA4, CHGB, LOC338749* or *TTN* fusion polypeptide in which the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide is fused at the N- or C-terminus to a string of histidine residues. Such fusion polypeptides can facilitate purification of recombinant *GP6, LAMA4, CHGB, LOC338749* or *TTN*. Expression vectors are commercially available that already encode a fusion moiety (e.g., a GST polypeptide), and a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid can be cloned into an expression vector such that the fusion moiety is linked in-frame to the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. Further, the fusion polypeptide can be a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide containing a heterologous signal sequence at its N-terminus. In certain host cells (e.g., mammalian host cells), expression, secretion, cellular internalization, and cellular localization of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide can be increased through use of a heterologous signal sequence. Fusion polypeptides can also include all or a part of a serum polypeptide (e.g., an IgG constant region or human serum albumin).

[0091] *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides or fragments thereof can be incorporated into pharmaceutical compositions and administered to a subject in vivo. Administration of these *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides can be used to affect the bioavailability of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* substrate and may effectively increase or decrease *GP6, LAMA4, CHGB, LOC338749* or *TTN* biological activity in a cell or effectively supplement dysfunctional or hyperactive *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. *GP6, LAMA4, CHGB, LOC338749* or *TTN* fusion polypeptides may be useful therapeutically for the treatment of disorders caused by, for example, (i) aberrant modification or mutation of a gene encoding a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide; (ii) mis-regulation of the *GP6, LAMA4, CHGB, LOC338749* or *TTN* gene; and (iii) aberrant post-translational modification of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. Also, *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides can be used as immunogens to produce anti-*GP6, LAMA4, CHGB, LOC338749* or *TTN* antibodies in a subject, to purify *GP6, LAMA4, CHGB, LOC338749* or *TTN* ligands or binding partners, and in screening assays to identify molecules which inhibit or enhance the interaction of *GP6, LAMA4, CHGB, LOC338749* or *TTN* with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* substrate. Preferably, said *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides are used in screening assays to identify molecules which inhibit the interaction of *GP6, LAMA4, CHGB, LOC338749* or *TTN*.

[0092] In addition, polypeptides can be chemically synthesized using techniques known in the art (See, *e.g.*, Creighton, 1983 Proteins. New York, N.Y.: W. H. Freeman and Company; and Hunkapiller *et al.*, (1984) Nature July 12 -18;310(5973):105-11). For example, a relative short polypeptide fragment can be synthesized by use of a peptide synthesizer. Furthermore, if desired, non-classical amino acids or chemical amino acid analogs can be introduced as a substitution or addition into the fragment sequence. Non-classical amino acids include, but are not limited to, to the D-isomers of the common amino acids, 2,4-diaminobutyric acid, a-amino isobutyric acid, 4-aminobutyric acid, Abu, 2-amino butyric acid, g-Abu, e-Ahx, 6-amino hexanoic acid, Aib, 2-amino isobutyric acid, 3-amino propionic acid, ornithine, norleucine, norvaline, hydroxyproline, sarcosine, citrulline, homocitrulline, cysteic acid, t-butylglycine, t-butylalanine, phenylglycine, cyclohexylalanine, b-alanine, fluoroamino acids, designer amino acids such as b-methyl amino acids, Ca-methyl amino acids, Na-methyl amino acids, and amino acid analogs in general. Furthermore, the amino acid can be D (dextrorotary) or L (levorotary).

[0093] Also included are polypeptide fragments which are differentially modified during or after translation, *e.g.*, by glycosylation, acetylation, phosphorylation, amidation, derivatization by known protecting/blocking groups, proteolytic cleavage, linkage to an antibody molecule or other cellular ligand, and the like. Any of numerous chemical modifications may be carried out by known techniques, including but not limited, to specific chemical cleavage by cyanogen bromide, trypsin, chymotrypsin,

papain, V8 protease, $NaBH_4$; acetylation, formylation, oxidation, reduction; metabolic synthesis in the presence of tunicamycin; and the like.

[0094] Additional post-translational modifications include, for example, N-linked or O-linked carbohydrate chains, processing of N-terminal or C-terminal ends), attachment of chemical moieties to the amino acid backbone, chemical modifications of N-linked or O-linked carbohydrate chains, and addition or deletion of an N-terminal methionine residue as a result of prokaryotic host cell expression. The polypeptide fragments may also be modified with a detectable label, such as an enzymatic, fluorescent, isotopic or affinity label to allow for detection and isolation of the polypeptide.

[0095] Also provided are chemically modified polypeptide derivatives that may provide additional advantages such as increased solubility, stability and circulating time of the polypeptide, or decreased immunogenicity. See U.S. Pat. No: 4,179,337. The chemical moieties for derivitization may be selected from water soluble polymers such as polyethylene glycol, ethylene glycol/propylene glycol copolymers, carboxymethylcellulose, dextran, polyvinyl alcohol and the like. The polypeptides may be modified at random positions within the molecule, or at predetermined positions within the molecule and may include one, two, three or more attached chemical moieties.

[0096] The polymer may be of any molecular weight, and may be branched or unbranched. For polyethylene glycol, the molecular weight is between about 1 kDa and about 100 kDa (the term "about" indicating that in preparations of polyethylene glycol, some molecules will weigh more, some less, than the stated molecular weight) for ease in handling and manufacturing. Other sizes may be used, depending on the desired therapeutic profile (*e.g.*, the duration of sustained release desired, the effects, if any on biological activity, the ease in handling, the degree or lack of antigenicity and other known effects of the polyethylene glycol to a therapeutic protein or analog).

[0097] The polyethylene glycol molecules (or other chemical moieties) should be attached to the polypeptide with consideration of effects on functional or antigenic domains of the polypeptide. There are a number of attachment methods available to those skilled in the art, *e.g.*, EP 0 401 384, herein incorporated by reference (coupling PEG to G-CSF), see also Malik *et al.* (1992) Exp Hematol. September;20(8):1028-35, reporting pegylation of GM-CSF using tresyl chloride). For example, polyethylene glycol may be covalently bound through amino acid residues via a reactive group, such as, a free amino or carboxyl group. Reactive groups are those to which an activated polyethylene glycol molecule may be bound. The amino acid residues having a free amino group may include lysine residues and the N-terminal amino acid residues; those having a free carboxyl group may include aspartic acid residues, glutamic acid residues and the C-terminal amino acid residue. Sulfhydryl groups may also be used as a reactive group for attaching the polyethylene glycol molecules. A polymer sometimes is attached at an amino group, such as attachment at the N-terminus or lysine group.

[0098] One may specifically desire proteins chemically modified at the N-terminus. Using polyethylene glycol as an illustration of the present composition, one may select from a variety of polyethylene glycol molecules (by molecular weight, branching, and the like), the proportion of polyethylene glycol molecules to protein (polypeptide) molecules in the reaction mix, the type of pegylation reaction to be performed, and the method of obtaining the selected N-terminally pegylated protein. The method of obtaining the N-terminally pegylated preparation (i.e., separating this moiety from other monopegylated moieties if necessary) may be by purification of the N-terminally pegylated material from a population of pegylated protein molecules. Selective proteins chemically modified at the N-terminus may be accomplished by reductive alkylation, which exploits differential reactivity of different types of primary amino groups (lysine versus the N-terminal) available for derivatization in a particular protein. Under the appropriate reaction conditions, substantially selective derivatization of the protein at the N-terminus with a carbonyl group containing polymer is achieved.

Substantially Identical Nucleic Acids and Polypeptides

[0099] Nucleotide sequences and polypeptide sequences that are substantially identical to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence and the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide sequences encoded by those nucleotide sequences are included herein. The term "substantially identical" as used herein refers to two or more nucleic acids or polypeptides sharing one or more identical nucleotide sequences or polypeptide sequences, respectively. Included are nucleotide sequences or polypeptide sequences that are 55% or more, 60% or more, 65% or more, 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more (each often within a 1%, 2%, 3% or 4% variability) or more identical to the nucleotide sequences in SEQ ID NO: 1-11 or the encoded *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide amino acid sequences. One test for determining whether two nucleic acids are substantially identical is to determine the percent of identical nucleotide sequences or polypeptide sequences shared between the nucleic acids or polypeptides.

[0100] Calculations of sequence identity are often performed as follows. Sequences are aligned for optimal comparison purposes (*e.g.*, gaps can be introduced in one or both of a first and a second amino acid or nucleic acid sequence for optimal alignment and non-homologous sequences can be disregarded for comparison purposes). The length of a reference sequence aligned for comparison purposes is sometimes 30% or more, 40% or more, 50% or more, often 60% or more, and more often 70% or more, 80% or more, 90% or more, 90% or more, or 100% of the length of the reference sequence. The nucleotides or amino acids at corresponding nucleotide or polypeptide positions, respectively, are then compared among the two sequences. When a position in the first sequence is occupied by the same nucleotide or amino acid as the corresponding position in the second sequence, the nucleotides or amino

acids are deemed to be identical at that position. The percent identity between the two sequences is a function of the number of identical positions shared by the sequences, taking into account the number of gaps, and the length of each gap, introduced for optimal alignment of the two sequences.

[0101] Comparison of sequences and determination of percent identity between two sequences can be accomplished using a mathematical algorithm. Percent identity between two amino acid or nucleotide sequences can be determined using the algorithm of Meyers & Miller, *CABIOS 4:* 11-17 (1989), which has been incorporated into the ALIGN program (version 2.0), using a PAM120 weight residue table, a gap length penalty of 12 and a gap penalty of 4. Also, percent identity between two amino acid sequences can be determined using the Needleman & Wunsch, *J. Mol. Biol. 48:* 444-453 (1970) algorithm which has been incorporated into the GAP program in the GCG software package (available at the http address www.gcg.com), using either a Blossum 62 matrix or a PAM250 matrix, and a gap weight of 16, 14, 12, 10, 8, 6, or 4 and a length weight of 1, 2, 3, 4, 5, or 6. Percent identity between two nucleotide sequences can be determined using the GAP program in the GCG software package (available at http address www.gcg.com), using a NWSgapdna.CMP matrix and a gap weight of 40, 50, 60, 70, or 80 and a length weight of 1, 2, 3, 4, 5, or 6. A set of parameters often used is a Blossum 62 scoring matrix with a gap open penalty of 12, a gap extend penalty of 4, and a frameshift gap penalty of 5.

[0102] Another manner for determining if two nucleic acids are substantially identical is to assess whether a polynucleotide homologous to one nucleic acid will hybridize to the other nucleic acid under stringent conditions. As use herein, the term "stringent conditions" refers to conditions for hybridization and washing. Stringent conditions are known to those skilled in the art and can be found in *Current Protocols in Molecular Biology*, John Wiley & Sons, N.Y. , 6.3.1-6.3.6 (1989). Aqueous and non-aqueous methods are described in that reference and either can be used. An example of stringent hybridization conditions is hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 50°C. Another example of stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 55°C. A further example of stringent hybridization conditions is hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 60°C. Often, stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 65°C. More often, stringency conditions are 0.5M sodium phosphate, 7% SDS at 65°C, followed by one or more washes at 0.2X SSC, 1% SDS at 65°C.

[0103] An example of a substantially identical nucleotide sequence to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence is one that has a different nucleotide sequence but still encodes the same polypeptide sequence encoded by the *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide

sequence. Another example is a nucleotide sequence that encodes a polypeptide having a polypeptide sequence that is more than 70% or more identical to, sometimes 75% or more, 80% or more, or 85% or more identical to, and often 90% or more and 95% or more identical to a polypeptide sequence encoded by a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence.

[0104] *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequences and *GP6, LAMA4, CHGB, LOC338749* or *TTN* amino acid sequences can be used as "query sequences" to perform a search against public databases to identify other family members or related sequences, for example. Such searches can be performed using the NBLAST and XBLAST programs (version 2.0) of Altschul *et al., J. Mol. Biol. 215:* 403-10 (1990). BLAST nucleotide searches can be performed with the NBLAST program, score = 100, wordlength = 12 to obtain nucleotide sequences homologous to nucleotide sequences from SEQ ID NO: 1-11. BLAST polypeptide searches can be performed with the XBLAST program, score = 50, wordlength = 3 to obtain amino acid sequences homologous to polypeptides encoded by a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence. To obtain gapped alignments for comparison purposes, Gapped BLAST can be utilized as described in Altschul *et al., Nucleic Acids Res. 25(17):* 3389-3402 (1997). When utilizing BLAST and Gapped BLAST programs, default parameters of the respective programs (*e.g.*, XBLAST and NBLAST) can be used (*see* the http address www.ncbi.nlm.nih.gov).

[0105] A nucleic acid that is substantially identical to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence may include polymorphic sites at positions equivalent to those described herein when the sequences are aligned. For example, using the alignment procedures described herein, SNPs in a sequence substantially identical to a sequence in SEQ ID NO: 1-11 can be identified at nucleotide positions that match (*i.e.*, align) with nucleotides at SNP positions in the nucleotide sequence of SEQ ID NO: 1-11. Also, where a polymorphic variation results in an insertion or deletion, insertion or deletion of a nucleotide sequence from a reference sequence can change the relative positions of other polymorphic sites in the nucleotide sequence.

[0106] Substantially identical nucleotide and polypeptide sequences include those that are naturally occurring, such as allelic variants (same locus), splice variants, homologs (different locus), and orthologs (different organism) or can be non-naturally occurring. Non-naturally occurring variants can be generated by mutagenesis techniques, including those applied to polynucleotides, cells, or organisms. The variants can contain nucleotide substitutions, deletions, inversions and insertions. Variation can occur in either or both the coding and non-coding regions. The variations can produce both conservative and non-conservative amino acid substitutions (as compared in the encoded product). Orthologs, homologs, allelic variants, and splice variants can be identified using methods known in the art. These variants normally comprise a nucleotide sequence encoding a polypeptide that is 50% or more, about

55% or more, often about 70-75% or more, more often about 80-85% or more, and typically about 90-95% or more identical to the amino acid sequences of target polypeptides or a fragment thereof. Such nucleic acid molecules readily can be identified as being able to hybridize under stringent conditions to a nucleotide sequence in SEQ ID NO: 1-11 or a fragment thereof. Nucleic acid molecules corresponding to orthologs, homologs, and allelic variants of a nucleotide sequence in SEQ ID NO: 1-11 can be identified by mapping the sequence to the same chromosome or locus as the nucleotide sequence in SEQ ID NO: 1-11.

[0107] Also, substantially identical nucleotide sequences may include codons that are altered with respect to the naturally occurring sequence for enhancing expression of a target polypeptide in a particular expression system. For example, the nucleic acid can be one in which one or more codons are altered, and often 10% or more or 20% or more of the codons are altered for optimized expression in bacteria (*e.g., E. coli.*), yeast (*e.g., S. cervesiae*), human (*e.g.*, 293 cells), insect, or rodent (*e.g.*, hamster) cells.

Methods for Identifying Subjects at Risk of Breast Cancer and Breast Cancer Risk in a Subject

[0108] Methods for prognosing and diagnosing breast cancer in subjects are provided herein. These methods include detecting the presence or absence of one or more polymorphic variations associated with breast cancer in a nucleotide sequence set forth in SEQ ID NO: 1-5, or substantially identical sequence thereof, in a sample from a subject, where the presence of a polymorphic variant is indicative of a risk of breast cancer.

[0109] Thus, featured herein is a method for detecting a subject at risk of breast cancer or the risk of breast cancer in a subject, which comprises detecting the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence set forth in SEQ ID NO: 1-5 in a nucleic acid sample from a subject, where the nucleotide sequence comprises a polynucleotide sequence selected from the group consisting of: (a) a nucleotide sequence set forth in SEQ ID NO: 1-5; (b) a nucleotide sequence which encodes a polypeptide having an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5; (c) a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5 or a nucleotide sequence about 90% or more identical to the nucleotide sequence set forth in SEQ ID NO: 1-5; and (d) a fragment of a nucleotide sequence of (a), (b), or (c), often a fragment that includes a polymorphic site associated with breast cancer; whereby the presence of the polymorphic variation is indicative of a risk of breast cancer in the subject. In certain embodiments, determining the presence of a combination of two or more polymorphic variants associated with breast cancer in one or more

nucleotide sequences of the sample is determined to identify a subject at risk of breast cancer and/or risk of breast cancer.

[0110] A risk of developing aggressive forms of breast cancer likely to metastasize or invade surrounding tissues (e.g., Stage IIIA, IIIB, and IV breast cancers), and subjects at risk of developing aggressive forms of breast cancer also may be identified by the methods described herein. These methods include collecting phenotype information from subjects having breast cancer, which includes the stage of progression of the breast cancer, and performing a secondary phenotype analysis to detect the presence or absence of one or more polymorphic variations associated with a particular stage form of breast cancer. Thus, detecting the presence or absence of one or more polymorphic variations in a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence associated with a late stage form of breast cancer often is prognostic and/or diagnostic of an aggressive form of the cancer.

[0111] Results from prognostic tests may be combined with other test results to diagnose breast cancer. For example, prognostic results may be gathered, a patient sample may be ordered based on a determined predisposition to breast cancer, the patient sample is analyzed, and the results of the analysis may be utilized to diagnose breast cancer. Also breast cancer diagnostic methods can be developed from studies used to generate prognostic/diagnostic methods in which populations are stratified into subpopulations having different progressions of breast cancer. In another embodiment, prognostic results may be gathered; a patient's risk factors for developing breast cancer analyzed (*e.g.*, age, race, family history, age of first menstrual cycle, age at birth of first child); and a patient sample may be ordered based on a determined predisposition to breast cancer. In an alternative embodiment, the results from predisposition analyses described herein may be combined with other test results indicative of breast cancer, which were previously, concurrently, or subsequently gathered with respect to the predisposition testing. In these embodiments, the combination of the prognostic test results with other test results can be probative of breast cancer, and the combination can be utilized as a breast cancer diagnostic. The results of any test indicative of breast cancer known in the art may be combined with the methods described herein. Examples of such tests are mammography (*e.g.*, a more frequent and/or earlier mammography regimen may be prescribed); breast biopsy and optionally a biopsy from another tissue; breast ultrasound and optionally an ultrasound analysis of another tissue; breast magnetic resonance imaging (MRI) and optionally an MRI analysis of another tissue; electrical impedance (T-scan) analysis of breast and optionally of another tissue; ductal lavage; nuclear medicine analysis (*e.g.*, scintimammography); *BRCA1* and/or *BRCA2* sequence analysis results; and thermal imaging of the breast and optionally of another tissue. Testing may be performed on tissue other than breast to diagnose the occurrence of metastasis (*e.g.*, testing of the lymph node).

[0112] Risk of breast cancer sometimes is expressed as a probability, such as an odds ratio, percentage, or risk factor. The risk is based upon the presence or absence of one or more polymorphic variants described herein, and also may be based in part upon phenotypic traits of the individual being tested. Methods for calculating predispositions based upon patient data are well known (*see, e.g.*, Agresti, *Categorical Data Analysis*, 2nd Ed. 2002. Wiley). Allelotyping and genotyping analyses may be carried out in populations other than those exemplified herein to enhance the predictive power of the prognostic method. These further analyses are executed in view of the exemplified procedures described herein, and may be based upon the same polymorphic variations or additional polymorphic variations. Risk determinations for breast cancer are useful in a variety of applications. In one embodiment, breast cancer risk determinations are used by clinicians to direct appropriate detection, preventative and treatment procedures to subjects who most require these. In another embodiment, breast cancer risk determinations are used by health insurers for preparing actuarial tables and for calculating insurance premiums.

[0113] The nucleic acid sample typically is isolated from a biological sample obtained from a subject. For example, nucleic acid can be isolated from blood, saliva, sputum, urine, cell scrapings, and biopsy tissue. The nucleic acid sample can be isolated from a biological sample using standard techniques, such as the technique described in Example 2. As used herein, the term "subject" refers primarily to humans but also refers to other mammals such as dogs, cats, and ungulates (*e.g.*, cattle, sheep, and swine). Subjects also include avians (*e.g.*, chickens and turkeys), reptiles, and fish (*e.g.*, salmon), as embodiments described herein can be adapted to nucleic acid samples isolated from any of these organisms. The nucleic acid sample may be isolated from the subject and then directly utilized in a method for determining the presence of a polymorphic variant, or alternatively, the sample may be isolated and then stored (*e.g.*, frozen) for a period of time before being subjected to analysis.

[0114] The presence or absence of a polymorphic variant is determined using one or both chromosomal complements represented in the nucleic acid sample. Determining the presence or absence of a polymorphic variant in both chromosomal complements represented in a nucleic acid sample from a subject having a copy of each chromosome is useful for determining the zygosity of an individual for the polymorphic variant (*i.e.*, whether the individual is homozygous or heterozygous for the polymorphic variant). Any oligonucleotide-based diagnostic may be utilized to determine whether a sample includes the presence or absence of a polymorphic variant in a sample. For example, primer extension methods, ligase sequence determination methods (*e.g.*, U.S. Pat. Nos. 5,679,524 and 5,952,174, and WO 01/27326), mismatch sequence determination methods (*e.g.*, U.S. Pat. Nos. 5,851,770; 5,958,692; 6,110,684; and 6,183,958), microarray sequence determination methods, restriction fragment length polymorphism (RFLP), single strand conformation polymorphism detection (SSCP) (*e.g.*, U.S. Pat. Nos.

5,891,625 and 6,013,499), PCR-based assays (*e.g.*, TAQMAN® PCR System (Applied Biosystems)), and nucleotide sequencing methods may be used.

[0115] Oligonucleotide extension methods typically involve providing a pair of oligonucleotide primers in a polymerase chain reaction (PCR) or in other nucleic acid amplification methods for the purpose of amplifying a region from the nucleic acid sample that comprises the polymorphic variation. One oligonucleotide primer is complementary to a region 3' of the polymorphism and the other is complementary to a region 5' of the polymorphism. A PCR primer pair may be used in methods disclosed in U.S. Pat. Nos. 4,683,195; 4,683,202, 4,965,188; 5,656,493; 5,998,143; 6,140,054; WO 01/27327; and WO 01/27329 for example. PCR primer pairs may also be used in any commercially available machines that perform PCR, such as any of the GENEAMP® Systems available from Applied Biosystems. Also, those of ordinary skill in the art will be able to design oligonucleotide primers based upon a nucleotide sequence set forth in SEQ ID NO: 1-5 without undue experimentation using knowledge readily available in the art.

[0116] Also provided is an extension oligonucleotide that hybridizes to the amplified fragment adjacent to the polymorphic variation. As used herein, the term "adjacent" refers to the 3' end of the extension oligonucleotide being often 1 nucleotide from the 5' end of the polymorphic site, and sometimes 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides from the 5' end of the polymorphic site, in the nucleic acid when the extension oligonucleotide is hybridized to the nucleic acid. The extension oligonucleotide then is extended by one or more nucleotides, and the number and/or type of nucleotides that are added to the extension oligonucleotide determine whether the polymorphic variant is present. Oligonucleotide extension methods are disclosed, for example, in U.S. Pat. Nos. 4,656,127; 4,851,331; 5,679,524; 5,834,189; 5,876,934; 5,908,755; 5,912,118; 5,976,802; 5,981,186; 6,004,744; 6,013,431; 6,017,702; 6,046,005; 6,087,095; 6,210,891; and WO 01/20039. Oligonucleotide extension methods using mass spectrometry are described, for example, in U.S. Pat. Nos. 5,547,835; 5,605,798; 5,691,141; 5,849,542; 5,869,242; 5,928,906; 6,043,031; and 6,194,144, and a method often utilized is described herein in Example 2. Multiple extension oligonucleotides may be utilized in one reaction, which is referred to herein as "multiplexing."

[0117] A microarray can be utilized for determining whether a polymorphic variant is present or absent in a nucleic acid sample. A microarray may include any oligonucleotides described herein, and methods for making and using oligonucleotide microarrays suitable for diagnostic use are disclosed in U.S. Pat. Nos. 5,492,806; 5,525,464; 5,589,330; 5,695,940; 5,849,483; 6,018,041; 6,045,996; 6,136,541; 6,142,681; 6,156,501; 6,197,506; 6,223,127; 6,225,625; 6,229,911; 6,239,273; WO 00/52625; WO 01/25485; and WO 01/29259. The microarray typically comprises a solid support and the oligonucleotides may be linked to this solid support by covalent bonds or by non-covalent interactions.

The oligonucleotides may also be linked to the solid support directly or by a spacer molecule. A microarray may comprise one or more oligonucleotides complementary to a polymorphic site set forth in SEQ ID NO: 1-5 or below.

[0118] A kit also may be utilized for determining whether a polymorphic variant is present or absent in a nucleic acid sample. A kit often comprises one or more pairs of oligonucleotide primers useful for amplifying a fragment of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence or a substantially identical sequence thereof, where the fragment includes a polymorphic site. The kit sometimes comprises a polymerizing agent, for example, a thermostable nucleic acid polymerase such as one disclosed in U.S. Pat. Nos. 4,889,818 or 6,077,664. Also, the kit often comprises an elongation oligonucleotide that hybridizes to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence in a nucleic acid sample adjacent to the polymorphic site. Where the kit includes an elongation oligonucleotide, it also often comprises chain elongating nucleotides, such as dATP, dTTP, dGTP, dCTP, and dITP, including analogs of dATP, dTTP, dGTP, dCTP and dITP, provided that such analogs are substrates for a thermostable nucleic acid polymerase and can be incorporated into a nucleic acid chain elongated from the extension oligonucleotide. Along with chain elongating nucleotides would be one or more chain terminating nucleotides such as ddATP, ddTTP, ddGTP, ddCTP, and the like. In an embodiment, the kit comprises one or more oligonucleotide primer pairs, a polymerizing agent, chain elongating nucleotides, at least one elongation oligonucleotide, and one or more chain terminating nucleotides. Kits optionally include buffers, vials, microtiter plates, and instructions for use.

[0119] An individual identified as being at risk of breast cancer may be heterozygous or homozygous with respect to the allele associated with a higher risk of breast cancer. A subject homozygous for an allele associated with an increased risk of breast cancer is at a comparatively high risk of breast cancer, a subject heterozygous for an allele associated with an increased risk of breast cancer is at a comparatively intermediate risk of breast cancer, and a subject homozygous for an allele associated with a decreased risk of breast cancer is at a comparatively low risk of breast cancer. A genotype may be assessed for a complementary strand, such that the complementary nucleotide at a particular position is detected.

[0120] Also featured are methods for determining risk of breast cancer and/or identifying a subject at risk of breast cancer by contacting a polypeptide or protein encoded by a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence from a subject with an antibody that specifically binds to an epitope associated with increased risk of breast cancer in the polypeptide. In certain embodiments, the antibody specifically binds to an epitope that comprises a lysine at amino acid position 237 of SEQ ID NO: 12, a proline at amino acid position 413 of SEQ ID NO: 16 or a glutamine at amino acid position 63 of SEQ ID NO: 16.

Applications of Prognostic and Diagnostic Results to Pharmacogenomic Methods

[0121] Pharmacogenomics is a discipline that involves tailoring a treatment for a subject according to the subject's genotype. For example, based upon the outcome of a prognostic test described herein, a clinician or physician may target pertinent information and preventative or therapeutic treatments to a subject who would be benefited by the information or treatment and avoid directing such information and treatments to a subject who would not be benefited (*e.g.*, the treatment has no therapeutic effect and/or the subject experiences adverse side effects). As therapeutic approaches for breast cancer continue to evolve and improve, the goal of treatments for breast cancer related disorders is to intervene even before clinical signs (*e.g.*, identification of lump in the breast) first manifest. Thus, genetic markers associated with susceptibility to breast cancer prove useful for early diagnosis, prevention and treatment of breast cancer.

[0122] The following is an example of a pharmacogenomic embodiment. A particular treatment regimen can exert a differential effect depending upon the subject's genotype. Where a candidate therapeutic exhibits a significant interaction with a major allele and a comparatively weak interaction with a minor allele (*e.g.*, an order of magnitude or greater difference in the interaction), such a therapeutic typically would not be administered to a subject genotyped as being homozygous for the minor allele, and sometimes not administered to a subject genotyped as being heterozygous for the minor allele. In another example, where a candidate therapeutic is not significantly toxic when administered to subjects who are homozygous for a major allele but is comparatively toxic when administered to subjects heterozygous or homozygous for a minor allele, the candidate therapeutic is not typically administered to subjects who are genotyped as being heterozygous or homozygous with respect to the minor allele.

[0123] The methods described herein are applicable to pharmacogenomic methods for detecting, preventing, alleviating and/or treating breast cancer. For example, a nucleic acid sample from an individual may be subjected to a genetic test described herein. Where one or more polymorphic variations associated with increased risk of breast cancer are identified in a subject, information for detecting, preventing or treating breast cancer and/or one or more breast cancer detection, prevention and/or treatment regimens then may be directed to and/or prescribed to that subject.

[0124] In certain embodiments, a detection, prevenative and/or treatment regimen is specifically prescribed and/or administered to individuals who will most benefit from it based upon their risk of developing breast cancer assessed by the methods described herein. Thus, provided are methods for identifying a subject at risk of breast cancer and then prescribing a detection, therapeutic or preventative regimen to individuals identified as being at risk of breast cancer. Thus, certain embodiments are directed to methods for treating breast cancer in a subject, reducing risk of breast cancer in a subject, or early detection of breast cancer in a subject, which comprise: detecting the presence or absence of a

polymorphic variant associated with breast cancer in a nucleic acid sample from a subject, where the nucleotide sequence comprises a polynucleotide sequence selected from the group consisting of: (a) a nucleotide sequence set forth in SEQ ID NO: 1-5; (b) a nucleotide sequence which encodes a polypeptide having an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5; (c) a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5 or a nucleotide sequence about 90% or more identical to the nucleotide sequence set forth in SEQ ID NO: 1-5; and (d) a fragment of a nucleotide sequence of (a), (b), or (c), sometimes comprising a polymorphic site associated with breast cancer; and prescribing or administering a breast cancer treatment regimen, preventative regimen and/or detection regimen to a subject from whom the sample originated where the presence of one or more polymorphic variations associated with breast cancer are detected in the nucleotide sequence. In these methods, genetic results may be utilized in combination with other test results to diagnose breast cancer as described above. Other test results include but are not limited to mammography results, imaging results, biopsy results and results from *BRCA1* or *BRAC2* test results, as described above.

[0125] Detection regimens include one or more mammography procedures, a regular mammography regimen (*e.g.*, once a year, or once every six, four, three or two months); an early mammography regimen (*e.g.*, mammography tests are performed beginning at age 25, 30, or 35); one or more biopsy procedures (*e.g.*, a regular biopsy regimen beginning at age 40); breast biopsy and biopsy from other tissue; breast ultrasound and optionally ultrasound analysis of another tissue; breast magnetic resonance imaging (MRI) and optionally MRI analysis of another tissue; electrical impedance (T-scan) analysis of breast and optionally another tissue; ductal lavage; nuclear medicine analysis (*e.g.*, scintimammography); *BRCA1* and/or *BRCA2* sequence analysis results; and/or thermal imaging of the breast and optionally another tissue.

[0126] Treatments sometimes are preventative (*e.g.*, is prescribed or administered to reduce the probability that a breast cancer associated condition arises or progresses), sometimes are therapeutic, and sometimes delay, alleviate or halt the progression of breast cancer. Any known preventative or therapeutic treatment for alleviating or preventing the occurrence of breast cancer is prescribed and/or administered. For example, certain preventative treatments often are prescribed to subjects having a predisposition to breast cancer and where the subject is not diagnosed with breast cancer or is diagnosed as having symptoms indicative of early stage breast cancer (*e.g.*, stage I). For subjects not diagnosed as having breast cancer, any preventative treatments known in the art can be prescribed and administered, which include selective hormone receptor modulators (*e.g.*, selective estrogen receptor modulators (SERMs) such as tamoxifen, reloxifene, and toremifene); compositions that prevent production of hormones (*e.g.*, aramotase inhibitors that prevent the production of estrogen in the adrenal gland, such as

exemestane, letrozole, anastrozol, groserelin, and megestrol); other hormonal treatments (*e.g.*, goserelin acetate and fulvestrant); biologic response modifiers such as antibodies (*e.g.*, trastuzumab (herceptin/HER2)); surgery (*e.g.*, lumpectomy and mastectomy); drugs that delay or halt metastasis (*e.g.*, pamidronate disodium); and alternative/complementary medicine (*e.g.*, acupuncture, acupressure, moxibustion, qi gong, reiki, ayurveda, vitamins, minerals, and herbs (*e.g.*, astragalus root, burdock root, garlic, green tea, and licorice root)).

[0127] The use of breast cancer treatments are well known in the art, and include surgery, chemotherapy and/or radiation therapy. Any of the treatments may be used in combination to treat or prevent breast cancer (*e.g.*, surgery followed by radiation therapy or chemotherapy). Examples of chemotherapy combinations used to treat breast cancer include: cyclophosphamide (Cytoxan), methotrexate (Amethopterin, Mexate, Folex), and fluorouracil (Fluorouracil, 5-Fu, Adrucil), which is referred to as CMF; cyclophosphamide, doxorubicin (Adriamycin), and fluorouracil, which is referred to as CAF; and doxorubicin (Adriamycin) and cyclophosphamide, which is referred to as AC.

[0128] As breast cancer preventative and treatment information can be specifically targeted to subjects in need thereof (*e.g.*, those at risk of developing breast cancer or those that have early signs of breast cancer), provided herein is a method for preventing or reducing the risk of developing breast cancer in a subject, which comprises: (a) detecting the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence in a nucleic acid sample from a subject; (b) identifying a subject with a predisposition to breast cancer, whereby the presence of the polymorphic variation is indicative of a predisposition to breast cancer in the subject; and (c) if such a predisposition is identified, providing the subject with information about methods or products to prevent or reduce breast cancer or to delay the onset of breast cancer. Also provided is a method of targeting information or advertising to a subpopulation of a human population based on the subpopulation being genetically predisposed to a disease or condition, which comprises: (a) detecting the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence in a nucleic acid sample from a subject; (b) identifying the subpopulation of subjects in which the polymorphic variation is associated with breast cancer; and (c) providing information only to the subpopulation of subjects about a particular product which may be obtained and consumed or applied by the subject to help prevent or delay onset of the disease or condition.

[0129] Pharmacogenomics methods also may be used to analyze and predict a response to a breast cancer treatment or a drug. For example, if pharmacogenomics analysis indicates a likelihood that an individual will respond positively to a breast cancer treatment with a particular drug, the drug may be administered to the individual. Conversely, if the analysis indicates that an individual is likely to respond negatively to treatment with a particular drug, an alternative course of treatment may be prescribed. A

negative response may be defined as either the absence of an efficacious response or the presence of toxic side effects. The response to a therapeutic treatment can be predicted in a background study in which subjects in any of the following populations are genotyped: a population that responds favorably to a treatment regimen, a population that does not respond significantly to a treatment regimen, and a population that responds adversely to a treatment regiment (*e.g.*, exhibits one or more side effects). These populations are provided as examples and other populations and subpopulations may be analyzed. Based upon the results of these analyses, a subject is genotyped to predict whether he or she will respond favorably to a treatment regimen, not respond significantly to a treatment regimen, or respond adversely to a treatment regimen.

[0130] The methods described herein also are applicable to clinical drug trials. One or more polymorphic variants indicative of response to an agent for treating breast cancer or to side effects to an agent for treating breast cancer may be identified using the methods described herein. Thereafter, potential participants in clinical trials of such an agent may be screened to identify those individuals most likely to respond favorably to the drug and exclude those likely to experience side effects. In that way, the effectiveness of drug treatment may be measured in individuals who respond positively to the drug, without lowering the measurement as a result of the inclusion of individuals who are unlikely to respond positively in the study and without risking undesirable safety problems. In certain embodiments, the agent for treating breast cancer described herein targets *GP6, LAMA4, CHGB, LOC338749* or *TTN* or a target in the *GP6, LAMA4, CHGB, LOC338749* or *TTN* pathway.

[0131] Thus, another embodiment is a method of selecting an individual for inclusion in a clinical trial of a treatment or drug comprising the steps of: (a) obtaining a nucleic acid sample from an individual; (b) determining the identity of a polymorphic variation which is associated with a positive response to the treatment or the drug, or at least one polymorphic variation which is associated with a negative response to the treatment or the drug in the nucleic acid sample, and (c) including the individual in the clinical trial if the nucleic acid sample contains said polymorphic variation associated with a positive response to the treatment or the drug or if the nucleic acid sample lacks said polymorphic variation associated with a negative response to the treatment or the drug. In addition, the methods for selecting an individual for inclusion in a clinical trial of a treatment or drug encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination. The polymorphic variation may be in a sequence selected individually or in any combination from the group consisting of (i) a polynucleotide sequence set forth in SEQ ID NO: 1-5; (ii) a polynucleotide sequence that is 90% or more identical to a nucleotide sequence set forth in SEQ ID NO: 1-5; (iii) a polynucleotide sequence that encodes a polypeptide having an amino acid sequence identical to or 90% or more identical to an amino acid sequence encoded by a nucleotide sequence set forth in SEQ ID NO:

1-5; and (iv) a fragment of a polynucleotide sequence of (i), (ii), or (iii) comprising the polymorphic site. The including step (c) optionally comprises administering the drug or the treatment to the individual if the nucleic acid sample contains the polymorphic variation associated with a positive response to the treatment or the drug and the nucleic acid sample lacks said biallelic marker associated with a negative response to the treatment or the drug.

[0132] Also provided herein is a method of partnering between a diagnostic/prognostic testing provider and a provider of a consumable product, which comprises: (a) the diagnostic/prognostic testing provider detects the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence in a nucleic acid sample from a subject; (b) the diagnostic/prognostic testing provider identifies the subpopulation of subjects in which the polymorphic variation is associated with breast cancer; (c) the diagnostic/prognostic testing provider forwards information to the subpopulation of subjects about a particular product which may be obtained and consumed or applied by the subject to help prevent or delay onset of the disease or condition; and (d) the provider of a consumable product forwards to the diagnostic test provider a fee every time the diagnostic/prognostic test provider forwards information to the subject as set forth in step (c) above.

Compositions Comprising Breast Cancer-Directed Molecules

[0133] Featured herein is a composition comprising a breast cancer cell and one or more molecules specifically directed and targeted to a nucleic acid comprising a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence or a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. Such directed molecules include, but are not limited to, a compound that binds to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid or a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide; a RNAi or siRNA molecule having a strand complementary to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence; an antisense nucleic acid complementary to an RNA encoded by a *GP6, LAMA4, CHGB, LOC338749* or *TTN* DNA sequence; a ribozyme that hybridizes to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence; a nucleic acid aptamer that specifically binds a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide; and an antibody that specifically binds to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or binds to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid. In certain embodiments, the antibody specifically binds to an epitope that comprises a a lysine at amino acid position 237 of SEQ ID NO: 12, a proline at amino acid position 413 of SEQ ID NO: 16 or a glutamine at amino acid position 63 of SEQ ID NO: 16. In specific embodiments, the breast cancer directed molecule interacts with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid or polypeptide variant associated with breast cancer. In other embodiments, the breast cancer directed molecule interacts with a polypeptide involved in the *GP6, LAMA4, CHGB, LOC338749* or *TTN* signal pathway,

or a nucleic acid encoding such a polypeptide. Polypeptides involved in the *GP6, LAMA4, CHGB, LOC338749* or *TTN* signal pathway are discussed herein.

[0134] Compositions sometimes include an adjuvant known to stimulate an immune response, and in certain embodiments, an adjuvant that stimulates a T-cell lymphocyte response. Adjuvants are known, including but not limited to an aluminum adjuvant (*e.g.*, aluminum hydroxide); a cytokine adjuvant or adjuvant that stimulates a cytokine response (*e.g.*, interleukin (IL)-12 and/or γ-interferon cytokines); a Freund-type mineral oil adjuvant emulsion (*e.g.*, Freund's complete or incomplete adjuvant); a synthetic lipoid compound; a copolymer adjuvant (*e.g.*, TitreMax); a saponin; Quil A; a liposome; an oil-in-water emulsion (*e.g.*, an emulsion stabilized by Tween 80 and pluronic polyoxyethlene/polyoxypropylene block copolymer (Syntex Adjuvant Formulation); TitreMax; detoxified endotoxin (MPL) and mycobacterial cell wall components (TDW, CWS) in 2% squalene (Ribi Adjuvant System)); a muramyl dipeptide; an immune-stimulating complex (ISCOM, *e.g.*, an Ag-modified saponin/cholesterol micelle that forms stable cage-like structure); an aqueous phase adjuvant that does not have a depot effect (*e.g.*, Gerbu adjuvant); a carbohydrate polymer (*e.g.*, AdjuPrime); L-tyrosine; a manide-oleate compound (*e.g.*, Montanide); an ethylene-vinyl acetate copolymer (*e.g.*, Elvax 40W1,2); or lipid A, for example. Such compositions are useful for generating an immune response against a breast cancer directed molecule (*e.g.*, an HLA-binding subsequence within a polypeptide encoded by a nucleotide sequence in SEQ ID NO: 1-5). In such methods, a peptide having an amino acid subsequence of a polypeptide encoded by a nucleotide sequence in SEQ ID NO: 1-5 is delivered to a subject, where the subsequence binds to an HLA molecule and induces a CTL lymphocyte response. The peptide sometimes is delivered to the subject as an isolated peptide or as a minigene in a plasmid that encodes the peptide. Methods for identifying HLA-binding subsequences in such polypeptides are known (*see e.g.*, publication WO02/20616 and PCT application US98/01373 for methods of identifying such sequences).

[0135] The breast cancer cell may be in a group of breast cancer cells and/or other types of cells cultured *in vitro* or in a tissue having breast cancer cells (*e.g.*, a melanocytic lesion) maintained *in vitro* or present in an animal *in vivo* (*e.g.*, a rat, mouse, ape or human). In certain embodiments, a composition comprises a component from a breast cancer cell or from a subject having a breast cancer cell instead of the breast cancer cell or in addition to the breast cancer cell, where the component sometimes is a nucleic acid molecule (*e.g.*, genomic DNA), a protein mixture or isolated protein, for example. The aforementioned compositions have utility in diagnostic, prognostic and pharmacogenomic methods described previously and in breast cancer therapeutics described hereafter. Certain breast cancer molecules are described in greater detail below.

Compounds

[0136] Compounds can be obtained using any of the numerous approaches in combinatorial library methods known in the art, including: biological libraries; peptoid libraries (libraries of molecules having the functionalities of peptides, but with a novel, non-peptide backbone which are resistant to enzymatic degradation but which nevertheless remain bioactive (see, *e.g.*, Zuckermann *et al.*, J. Med. Chem.37: 2678-85 (1994)); spatially addressable parallel solid phase or solution phase libraries; synthetic library methods requiring deconvolution; "one-bead one-compound" library methods; and synthetic library methods using affinity chromatography selection. Biological library and peptoid library approaches are typically limited to peptide libraries, while the other approaches are applicable to peptide, non-peptide oligomer or small molecule libraries of compounds (Lam, Anticancer Drug Des. 12: 145, (1997)). Examples of methods for synthesizing molecular libraries are described, for example, in DeWitt *et al.*, Proc. Natl. Acad. Sci. U.S.A. 90: 6909 (1993); Erb *et al.*, Proc. Natl. Acad. Sci. USA 91: 11422 (1994); Zuckermann *et al.*, J. Med. Chem. 37: 2678 (1994); Cho *et al.*, Science 261: 1303 (1993); Carrell *et al.*, Angew. Chem. Int. Ed. Engl. 33: 2059 (1994); Carell *et al.*, Angew. Chem. Int. Ed. Engl. 33: 2061 (1994); and in Gallop *et al.*, J. Med. Chem. 37: 1233 (1994).

[0137] Libraries of compounds may be presented in solution (*e.g.*, Houghten, Biotechniques 13: 412-421 (1992)), or on beads (Lam, Nature 354: 82-84 (1991)), chips (Fodor, Nature 364: 555-556 (1993)), bacteria or spores (Ladner, United States Patent No. 5,223,409), plasmids (Cull *et al.*, Proc. Natl. Acad. Sci. USA 89: 1865-1869 (1992)) or on phage (Scott and Smith, Science 249: 386-390 (1990); Devlin, Science 249: 404-406 (1990); Cwirla *et al.*, Proc. Natl. Acad. Sci. 87: 6378-6382 (1990); Felici, J. Mol. Biol. 222: 301-310 (1991); Ladner supra.).

[0138] A compound sometimes alters expression and sometimes alters activity of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide and may be a small molecule. Small molecules include, but are not limited to, peptides, peptidomimetics (*e.g.*, peptoids), amino acids, amino acid analogs, polynucleotides, polynucleotide analogs, nucleotides, nucleotide analogs, organic or inorganic compounds (i.e., including heteroorganic and organometallic compounds) having a molecular weight less than about 10,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 5,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 1,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 500 grams per mole, and salts, esters, and other pharmaceutically acceptable forms of such compounds.

Antisense Nucleic Acid Molecules, Ribozymes, RNAi, siRNA and Modified Nucleic Acid Molecules

[0139] An "antisense" nucleic acid refers to a nucleotide sequence complementary to a "sense" nucleic acid encoding a polypeptide, *e.g.*, complementary to the coding strand of a double-stranded cDNA molecule or complementary to an mRNA sequence. The antisense nucleic acid can be complementary to an entire coding strand in SEQ ID NO: 1-11, or to a portion thereof or a substantially identical sequence thereof. In another embodiment, the antisense nucleic acid molecule is antisense to a "noncoding region" of the coding strand of a nucleotide sequence in SEQ ID NO: 1-11 (*e.g.*, 5' and 3' untranslated regions).

[0140] An antisense nucleic acid can be designed such that it is complementary to the entire coding region of an mRNA encoded by a nucleotide sequence in SEQ ID NO: 1-5 (e.g., SEQ ID NO: 6-11), and often the antisense nucleic acid is an oligonucleotide antisense to only a portion of a coding or noncoding region of the mRNA. For example, the antisense oligonucleotide can be complementary to the region surrounding the translation start site of the mRNA, *e.g.*, between the -10 and +10 regions of the target gene nucleotide sequence of interest. An antisense oligonucleotide can be, for example, about 7, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, or more nucleotides in length. The antisense nucleic acids, which include the ribozymes described hereafter, can be designed to target a nucleotide sequence in SEQ ID NO: 1-11, often a variant associated with breast cancer, or a substantially identical sequence thereof. Among the variants, minor alleles and major alleles can be targeted, and those associated with a higher risk of breast cancer are often designed, tested, and administered to subjects.

[0141] An antisense nucleic acid can be constructed using chemical synthesis and enzymatic ligation reactions using standard procedures. For example, an antisense nucleic acid (*e.g.*, an antisense oligonucleotide) can be chemically synthesized using naturally occurring nucleotides or variously modified nucleotides designed to increase the biological stability of the molecules or to increase the physical stability of the duplex formed between the antisense and sense nucleic acids, *e.g.*, phosphorothioate derivatives and acridine substituted nucleotides can be used. Antisense nucleic acid also can be produced biologically using an expression vector into which a nucleic acid has been subcloned in an antisense orientation (i.e., RNA transcribed from the inserted nucleic acid will be of an antisense orientation to a target nucleic acid of interest, described further in the following subsection).

[0142] When utilized as therapeutics, antisense nucleic acids typically are administered to a subject (*e.g.*, by direct injection at a tissue site) or generated in situ such that they hybridize with or bind to cellular mRNA and/or genomic DNA encoding a polypeptide and thereby inhibit expression of the polypeptide, for example, by inhibiting transcription and/or translation. Alternatively, antisense nucleic acid molecules can be modified to target selected cells and then are administered systemically. For

systemic administration, antisense molecules can be modified such that they specifically bind to receptors or antigens expressed on a selected cell surface, for example, by linking antisense nucleic acid molecules to peptides or antibodies which bind to cell surface receptors or antigens. Antisense nucleic acid molecules can also be delivered to cells using the vectors described herein. Sufficient intracellular concentrations of antisense molecules are achieved by incorporating a strong promoter, such as a pol II or pol III promoter, in the vector construct.

[0143] Antisense nucleic acid molecules sometimes are *-anomeric nucleic acid molecules. An *-anomeric nucleic acid molecule forms specific double-stranded hybrids with complementary RNA in which, contrary to the usual *-units, the strands run parallel to each other (Gaultier *et al.*, Nucleic Acids. Res. 15: 6625-6641 (1987)). Antisense nucleic acid molecules can also comprise a 2'-o-methylribonucleotide (Inoue *et al.*, Nucleic Acids Res. 15: 6131-6148 (1987)) or a chimeric RNA-DNA analogue (Inoue *et al.*, FEBS Lett. 215: 327-330 (1987)). Antisense nucleic acids sometimes are composed of DNA or PNA or any other nucleic acid derivatives described previously.

[0144] In another embodiment, an antisense nucleic acid is a ribozyme. A ribozyme having specificity for a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence can include one or more sequences complementary to such a nucleotide sequence, and a sequence having a known catalytic region responsible for mRNA cleavage (see *e.g.*, U.S. Pat. No. 5,093,246 or Haselhoff and Gerlach, Nature 334: 585-591 (1988)). For example, a derivative of a Tetrahymena L-19 IVS RNA is sometimes utilized in which the nucleotide sequence of the active site is complementary to the nucleotide sequence to be cleaved in a mRNA (see *e.g.*, Cech *et al.* U.S. Patent No. 4,987,071; and Cech *et al.* U.S. Patent No. 5,116,742). Also, target mRNA sequences can be used to select a catalytic RNA having a specific ribonuclease activity from a pool of RNA molecules (see *e.g.*, Bartel & Szostak, Science 261: 1411-1418 (1993)).

[0145] Breast cancer directed molecules include in certain embodiments nucleic acids that can form triple helix structures with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence or a substantially identical sequence thereof, especially one that includes a regulatory region that controls expression of a polypeptide. Gene expression can be inhibited by targeting nucleotide sequences complementary to the regulatory region of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence or a substantially identical sequence (*e.g.*, promoter and/or enhancers) to form triple helical structures that prevent transcription of a gene in target cells (see *e.g.*, Helene, Anticancer Drug Des. 6(6): 569-84 (1991); Helene *et al.*, Ann. N.Y. Acad. Sci. 660: 27-36 (1992); and Maher, Bioassays 14(12): 807-15 (1992). Potential sequences that can be targeted for triple helix formation can be increased by creating a so-called "switchback" nucleic acid molecule. Switchback molecules are synthesized in an alternating 5'-3', 3'-5' manner, such that they base pair with first one strand of a duplex and then the

other, eliminating the necessity for a sizeable stretch of either purines or pyrimidines to be present on one strand of a duplex.

[0146] Breast cancer directed molecules include RNAi and siRNA nucleic acids. Gene expression may be inhibited by the introduction of double-stranded RNA (dsRNA), which induces potent and specific gene silencing, a phenomenon called RNA interference or RNAi. See, *e.g.*, Fire *et al.*, US Patent Number 6,506,559; Tuschl *et al.* PCT International Publication No. WO 01/75164; Kay *et al.* PCT International Publication No. WO 03/010180A1; or Bosher JM, Labouesse, Nat Cell Biol 2000 Feb;2(2):E31-6. This process has been improved by decreasing the size of the double-stranded RNA to 20-24 base pairs (to create small-interfering RNAs or siRNAs) that "switched off" genes in mammalian cells without initiating an acute phase response, i.e., a host defense mechanism that often results in cell death (see, *e.g.*, Caplen *et al.* Proc Natl Acad Sci U S A. 2001 Aug 14;98(17):9742-7 and Elbashir *et al.* Methods 2002 Feb;26(2):199-213). There is increasing evidence of post-transcriptional gene silencing by RNA interference (RNAi) for inhibiting targeted expression in mammalian cells at the mRNA level, in human cells. There is additional evidence of effective methods for inhibiting the proliferation and migration of tumor cells in human patients, and for inhibiting metastatic cancer development (see, *e.g.*, U.S. Patent Application No. US2001000993183; Caplen *et al.* Proc Natl Acad Sci U S A; and Abderrahmani *et al.* Mol Cell Biol 2001 Nov21(21):7256-67).

[0147] An "siRNA" or "RNAi" refers to a nucleic acid that forms a double stranded RNA and has the ability to reduce or inhibit expression of a gene or target gene when the siRNA is delivered to or expressed in the same cell as the gene or target gene. "siRNA" refers to short double-stranded RNA formed by the complementary strands. Complementary portions of the siRNA that hybridize to form the double stranded molecule often have substantial or complete identity to the target molecule sequence. In one embodiment, an siRNA refers to a nucleic acid that has substantial or complete identity to a target gene and forms a double stranded siRNA.

[0148] When designing the siRNA molecules, the targeted region often is selected from a given DNA sequence beginning 50 to 100 nucleotides downstream of the start codon. See, *e.g.*, Elbashir et al,. Methods 26:199-213 (2002). Initially, 5' or 3' UTRs and regions nearby the start codon were avoided assuming that UTR-binding proteins and/or translation initiation complexes may interfere with binding of the siRNP or RISC endonuclease complex. Sometimes regions of the target 23 nucleotides in length conforming to the sequence motif AA(N19)TT (N, an nucleotide), and regions with approximately 30% to 70% G/C-content (often about 50% G/C-content) often are selected. If no suitable sequences are found, the search often is extended using the motif NA(N21). The sequence of the sense siRNA sometimes corresponds to (N19) TT or N21 (position 3 to 23 of the 23-nt motif), respectively. In the latter case, the 3' end of the sense siRNA often is converted to TT. The rationale for this sequence

conversion is to generate a symmetric duplex with respect to the sequence composition of the sense and antisense 3' overhangs. The antisense siRNA is synthesized as the complement to position 1 to 21 of the 23-nt motif. Because position 1 of the 23-nt motif is not recognized sequence-specifically by the antisense siRNA, the 3'-most nucleotide residue of the antisense siRNA can be chosen deliberately. However, the penultimate nucleotide of the antisense siRNA (complementary to position 2 of the 23-nt motif) often is complementary to the targeted sequence. For simplifying chemical synthesis, TT often is utilized. siRNAs corresponding to the target motif NAR(N17)YNN, where R is purine (A,G) and Y is pyrimidine (C,U), often are selected. Respective 21 nucleotide sense and antisense siRNAs often begin with a purine nucleotide and can also be expressed from pol III expression vectors without a change in targeting site. Expression of RNAs from pol III promoters often is efficient when the first transcribed nucleotide is a purine.

[0149] The sequence of the siRNA can correspond to the full length target gene, or a subsequence thereof. Often, the siRNA is about 15 to about 50 nucleotides in length (*e.g.*, each complementary sequence of the double stranded siRNA is 15-50 nucleotides in length, and the double stranded siRNA is about 15-50 base pairs in length, sometimes about 20-30 nucleotides in length or about 20-25 nucleotides in length, *e.g.*, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides in length. The siRNA sometimes is about 21 nucleotides in length. Methods of using siRNA are well known in the art, and specific siRNA molecules may be purchased from a number of companies including Dharmacon Research, Inc.

[0150] Antisense, ribozyme, RNAi and siRNA nucleic acids can be altered to form modified nucleic acid molecules. The nucleic acids can be altered at base moieties, sugar moieties or phosphate backbone moieties to improve stability, hybridization, or solubility of the molecule. For example, the deoxyribose phosphate backbone of nucleic acid molecules can be modified to generate peptide nucleic acids (see Hyrup *et al.*, Bioorganic & Medicinal Chemistry 4 (1): 5-23 (1996)). As used herein, the terms "peptide nucleic acid" or "PNA" refers to a nucleic acid mimic such as a DNA mimic, in which the deoxyribose phosphate backbone is replaced by a pseudopeptide backbone and only the four natural nucleobases are retained. The neutral backbone of a PNA can allow for specific hybridization to DNA and RNA under conditions of low ionic strength. Synthesis of PNA oligomers can be performed using standard solid phase peptide synthesis protocols as described, for example, in Hyrup *et al.*, (1996) supra and Perry-O'Keefe *et al.*, Proc. Natl. Acad. Sci. 93: 14670-675 (1996).

[0151] PNA nucleic acids can be used in prognostic, diagnostic, and therapeutic applications. For example, PNAs can be used as antisense or antigene agents for sequence-specific modulation of gene expression by, for example, inducing transcription or translation arrest or inhibiting replication. PNA nucleic acid molecules can also be used in the analysis of single base pair mutations in a gene, (*e.g.*, by PNA-directed PCR clamping); as "artificial restriction enzymes" when used in combination with other

enzymes, (*e.g.*, S1 nucleases (Hyrup (1996) supra)); or as probes or primers for DNA sequencing or hybridization (Hyrup *et al.*, (1996) supra; Perry-O'Keefe supra).

[0152] In other embodiments, oligonucleotides may include other appended groups such as peptides (*e.g.*, for targeting host cell receptors *in vivo*), or agents facilitating transport across cell membranes (see *e.g.*, Letsinger *et al.*, Proc. Natl. Acad. Sci. USA 86: 6553-6556 (1989); Lemaitre *et al.*, Proc. Natl. Acad. Sci. USA 84: 648-652 (1987); PCT Publication No. W088/09810) or the blood-brain barrier (see, *e.g.*, PCT Publication No. W089/10134). In addition, oligonucleotides can be modified with hybridization-triggered cleavage agents (See, *e.g.*, Krol *et al.*, Bio-Techniques 6: 958-976 (1988)) or intercalating agents. (See, *e.g.*, Zon, Pharm. Res. 5: 539-549 (1988)). To this end, the oligonucleotide may be conjugated to another molecule, (*e.g.*, a peptide, hybridization triggered cross-linking agent, transport agent, or hybridization-triggered cleavage agent).

[0153] Also included herein are molecular beacon oligonucleotide primer and probe molecules having one or more regions complementary to a nucleotide sequence of SEQ ID NO: 1-11 or a substantially identical sequence thereof, two complementary regions one having a fluorophore and one a quencher such that the molecular beacon is useful for quantifying the presence of the nucleic acid in a sample. Molecular beacon nucleic acids are described, for example, in Lizardi *et al.*, U.S. Patent No. 5,854,033; Nazarenko *et al.*, U.S. Patent No. 5,866,336, and Livak *et al.*, U.S. Patent 5,876,930.

Antibodies

[0154] The term "antibody" as used herein refers to an immunoglobulin molecule or immunologically active portion thereof, i.e., an antigen-binding portion. Examples of immunologically active portions of immunoglobulin molecules include F(ab) and F(ab')$_2$ fragments which can be generated by treating the antibody with an enzyme such as pepsin. An antibody sometimes is a polyclonal, monoclonal, recombinant (*e.g.*, a chimeric or humanized), fully human, non-human (*e.g.*, murine), or a single chain antibody. An antibody may have effector function and can fix complement, and is sometimes coupled to a toxin or imaging agent.

[0155] A full-length polypeptide or antigenic peptide fragment encoded by a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleotide sequence can be used as an immunogen or can be used to identify antibodies made with other immunogens, *e.g.*, cells, membrane preparations, and the like. An antigenic peptide often includes at least 8 amino acid residues of the amino acid sequences encoded by a nucleotide sequence of SEQ ID NO: 1-11, or substantially identical sequence thereof, and encompasses an epitope. Antigenic peptides sometimes include 10 or more amino acids, 15 or more amino acids, 20 or more amino acids, or 30 or more amino acids. Hydrophilic and hydrophobic fragments of polypeptides sometimes are used as immunogens.

[0156] Epitopes encompassed by the antigenic peptide are regions located on the surface of the polypeptide (*e.g.*, hydrophilic regions) as well as regions with high antigenicity. For example, an Emini surface probability analysis of the human polypeptide sequence can be used to indicate the regions that have a particularly high probability of being localized to the surface of the polypeptide and are thus likely to constitute surface residues useful for targeting antibody production. The antibody may bind an epitope on any domain or region on polypeptides described herein.

[0157] Also, chimeric, humanized, and completely human antibodies are useful for applications which include repeated administration to subjects. Chimeric and humanized monoclonal antibodies, comprising both human and non-human portions, can be made using standard recombinant DNA techniques. Such chimeric and humanized monoclonal antibodies can be produced by recombinant DNA techniques known in the art, for example using methods described in Robinson et al International Application No. PCT/US86/02269; Akira, et al European Patent Application 184,187; Taniguchi, M., European Patent Application 171,496; Morrison et al European Patent Application 173,494; Neuberger et al PCT International Publication No. WO 86/01533; Cabilly et al U.S. Patent No. 4,816,567; Cabilly et al European Patent Application 125,023; Better *et al.*, Science 240: 1041-1043 (1988); Liu *et al.*, Proc. Natl. Acad. Sci. USA 84: 3439-3443 (1987); Liu *et al.*, J. Immunol. 139: 3521-3526 (1987); Sun *et al.*, Proc. Natl. Acad. Sci. USA 84: 214-218 (1987); Nishimura *et al.*, Canc. Res. 47: 999-1005 (1987); Wood *et al.*, Nature 314: 446-449 (1985); and Shaw *et al.*, J. Natl. Cancer Inst. 80: 1553-1559 (1988); Morrison, S. L., Science 229: 1202-1207 (1985); Oi *et al.*, BioTechniques 4: 214 (1986); Winter U.S. Patent 5,225,539; Jones *et al.*, Nature 321: 552-525 (1986); Verhoeyan *et al.*, Science 239: 1534; and Beidler *et al.*, J. Immunol. 141: 4053-4060 (1988).

[0158] Completely human antibodies are particularly desirable for therapeutic treatment of human patients. Such antibodies can be produced using transgenic mice that are incapable of expressing endogenous immunoglobulin heavy and light chains genes, but which can express human heavy and light chain genes. See, for example, Lonberg and Huszar, Int. Rev. Immunol. 13: 65-93 (1995); and U.S. Patent Nos. 5,625,126; 5,633,425; 5,569,825; 5,661,016; and 5,545,806. In addition, companies such as Abgenix, Inc. (Fremont, CA) and Medarex, Inc. (Princeton, NJ), can be engaged to provide human antibodies directed against a selected antigen using technology similar to that described above. Completely human antibodies that recognize a selected epitope also can be generated using a technique referred to as "guided selection." In this approach a selected non-human monoclonal antibody (*e.g.*, a murine antibody) is used to guide the selection of a completely human antibody recognizing the same epitope. This technology is described for example by Jespers *et al.*, Bio/Technology 12: 899-903 (1994).

[0159] Antibody can be a single chain antibody. A single chain antibody (scFV) can be engineered (see, *e.g.*, Colcher *et al.*, Ann. N Y Acad. Sci. 880: 263-80 (1999); and Reiter, Clin. Cancer Res. 2: 245-

52 (1996)). Single chain antibodies can be dimerized or multimerized to generate multivalent antibodies having specificities for different epitopes of the same target polypeptide.

[0160] Antibodies also may be selected or modified so that they exhibit reduced or no ability to bind an Fc receptor. For example, an antibody may be an isotype or subtype, fragment or other mutant, which does not support binding to an Fc receptor (*e.g.*, it has a mutagenized or deleted Fc receptor binding region).

[0161] Also, an antibody (or fragment thereof) may be conjugated to a therapeutic moiety such as a cytotoxin, a therapeutic agent or a radioactive metal ion. A cytotoxin or cytotoxic agent includes any agent that is detrimental to cells. Examples include taxol, cytochalasin B, gramicidin D, ethidium bromide, emetine, mitomycin, etoposide, tenoposide, vincristine, vinblastine, colchicin, doxorubicin, daunorubicin, dihydroxy anthracin dione, mitoxantrone, mithramycin, actinomycin D, 1 dehydrotestosterone, glucocorticoids, procaine, tetracaine, lidocaine, propranolol, and puromycin and analogs or homologs thereof. Therapeutic agents include, but are not limited to, antimetabolites (*e.g.*, methotrexate, 6-mercaptopurine, 6-thioguanine, cytarabine, 5-fluorouracil decarbazine), alkylating agents (*e.g.*, mechlorethamine, thiotepa chlorambucil, melphalan, carmustine (BCNU) and lomustine (CCNU), cyclophosphamide, busulfan, dibromomannitol, streptozotocin, mitomycin C, and cis-dichlorodiamine platinum (II) (DDP) cisplatin), anthracyclines (*e.g.*, daunorubicin (formerly daunomycin) and doxorubicin), antibiotics (*e.g.*, dactinomycin (formerly actinomycin), bleomycin, mithramycin, and anthramycin (AMC)), and anti-mitotic agents (*e.g.*, vincristine and vinblastine).

[0162] Antibody conjugates can be used for modifying a given biological response. For example, the drug moiety may be a protein or polypeptide possessing a desired biological activity. Such proteins may include, for example, a toxin such as abrin, ricin A, pseudomonas exotoxin, or diphtheria toxin; a polypeptide such as tumor necrosis factor, γ-interferon, α-interferon, nerve growth factor, platelet derived growth factor, tissue plasminogen activator; or, biological response modifiers such as, for example, lymphokines, interleukin-1 ("IL-1"), interleukin-2 ("IL-2"), interleukin-6 ("IL-6"), granulocyte macrophage colony stimulating factor ("GM-CSF"), granulocyte colony stimulating factor ("G-CSF"), or other growth factors. Also, an antibody can be conjugated to a second antibody to form an antibody heteroconjugate as described by Segal in U.S. Patent No. 4,676,980, for example.

[0163] An antibody (*e.g.*, monoclonal antibody) can be used to isolate target polypeptides by standard techniques, such as affinity chromatography or immunoprecipitation. Moreover, an antibody can be used to detect a target polypeptide (*e.g.*, in a cellular lysate or cell supernatant) in order to evaluate the abundance and pattern of expression of the polypeptide. Antibodies can be used diagnostically to monitor polypeptide levels in tissue as part of a clinical testing procedure, *e.g.*, to determine the efficacy of a given treatment regimen. Detection can be facilitated by coupling (i.e., physically linking) the

antibody to a detectable substance (i.e., antibody labeling). Examples of detectable substances include various enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, β-galactosidase, or acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin; examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include luciferase, luciferin, and aequorin, and examples of suitable radioactive material include ^{125}I, ^{131}I, ^{35}S or ^{3}H. Also, an antibody can be utilized as a test molecule for determining whether it can treat breast cancer, and as a therapeutic for administration to a subject for treating breast cancer.

[0164] An antibody can be made by immunizing with a purified antigen, or a fragment thereof, *e.g.*, a fragment described herein, a membrane associated antigen, tissues, *e.g.*, crude tissue preparations, whole cells, preferably living cells, lysed cells, or cell fractions.

[0165] Included herein are antibodies which bind only a native polypeptide, only denatured or otherwise non-native polypeptide, or which bind both, as well as those having linear or conformational epitopes. Conformational epitopes sometimes can be identified by selecting antibodies that bind to native but not denatured polypeptide. Also featured are antibodies that specifically bind to a polypeptide variant associated with breast cancer.

Screening Assays

[0166] Featured herein are methods for identifying a candidate therapeutic for treating breast cancer. The methods comprise contacting a test molecule with a target molecule in a system. A "target molecule" as used herein refers to a nucleic acid of SEQ ID NO: 1-11, a substantially identical nucleic acid thereof, or a fragment thereof, and an encoded polypeptide of the foregoing. The method also comprises determining the presence or absence of an interaction between the test molecule and the target molecule, where the presence of an interaction between the test molecule and the nucleic acid or polypeptide identifies the test molecule as a candidate breast cancer therapeutic. The interaction between the test molecule and the target molecule may be quantified.

[0167] Test molecules and candidate therapeutics include, but are not limited to, compounds, antisense nucleic acids, siRNA molecules, ribozymes, polypeptides or proteins encoded by a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acids, or a substantially identical sequence or fragment thereof, and immunotherapeutics (*e.g.*, antibodies and HLA-presented polypeptide fragments). A test molecule or candidate therapeutic may act as a modulator of target molecule concentration or target

molecule function in a system. A "modulator" may agonize (i.e., up-regulates) or antagonize (i.e., down-regulates) a target molecule concentration partially or completely in a system by affecting such cellular functions as DNA replication and/or DNA processing (*e.g.*, DNA methylation or DNA repair), RNA transcription and/or RNA processing (*e.g.*, removal of intronic sequences and/or translocation of spliced mRNA from the nucleus), polypeptide production (*e.g.*, translation of the polypeptide from mRNA), and/or polypeptide post-translational modification (*e.g.*, glycosylation, phosphorylation, and proteolysis of pro-polypeptides). A modulator may also agonize or antagonize a biological function of a target molecule partially or completely, where the function may include adopting a certain structural conformation, interacting with one or more binding partners, ligand binding, catalysis (*e.g.*, phosphorylation, dephosphorylation, hydrolysis, methylation, and isomerization), and an effect upon a cellular event (*e.g.*, effecting progression of breast cancer).

[0168] As used herein, the term "system" refers to a cell free *in vitro* environment and a cell-based environment such as a collection of cells, a tissue, an organ, or an organism. A system is "contacted" with a test molecule in a variety of manners, including adding molecules in solution and allowing them to interact with one another by diffusion, cell injection, and any administration routes in an animal. As used herein, the term "interaction" refers to an effect of a test molecule on test molecule, where the effect sometimes is binding between the test molecule and the target molecule, and sometimes is an observable change in cells, tissue, or organism.

[0169] There are many standard methods for detecting the presence or absence of an interaction between a test molecule and a target molecule. For example, titrametric, acidimetric, radiometric, NMR, monolayer, polarographic, spectrophotometric, fluorescent, and ESR assays probative of a target molecule interaction may be utilized.

[0170] In general, an interaction can be determined by labeling the test molecule and/or the *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule, where the label is covalently or non-covalently attached to the test molecule or *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule. The label is sometimes a radioactive molecule such as ^{125}I, ^{131}I, ^{35}S or ^{3}H, which can be detected by direct counting of radioemission or by scintillation counting. Also, enzymatic labels such as horseradish peroxidase, alkaline phosphatase, or luciferase may be utilized where the enzymatic label can be detected by determining conversion of an appropriate substrate to product. Also, presence or absence of an interaction can be determined without labeling. For example, a microphysiometer (e.g., Cytosensor) is an analytical instrument that measures the rate at which a cell acidifies its environment using a light-addressable potentiometric sensor (LAPS). Changes in this acidification rate can be used as an indication of an interaction between a test molecule and *GP6, LAMA4, CHGB, LOC338749* or *TTN* (McConnell, H. M. et al., Science 257: 1906-1912 (1992)).

[0171] In cell-based systems, cells typically include a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid or polypeptide or variants thereof and are often of mammalian origin, although the cell can be of any origin. Whole cells, cell homogenates, and cell fractions (e.g., cell membrane fractions) can be subjected to analysis. Where interactions between a test molecule with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or variant thereof are monitored, soluble and/or membrane bound forms of the polypeptide or variant may be utilized. Where membrane-bound forms of the polypeptide are used, it may be desirable to utilize a solubilizing agent. Examples of such solubilizing agents include non-ionic detergents such as n-octylglucoside, n-dodecylglucoside, n-dodecylmaltoside, octanoyl-N-methylglucamide, decanoyl-N-methylglucamide, Triton® X-100, Triton® X-114, Thesit®, Isotridecypoly(ethylene glycol ether)n, 3-[(3-cholamidopropyl)dimethylamminio]-1-propane sulfonate (CHAPS), 3-[(3-cholamidopropyl)dimethylamminio]-2-hydroxy-1-propane sulfonate (CHAPSO), or N-dodecyl-N,N-dimethyl-3-ammonio-1-propane sulfonate.

[0172] An interaction between two molecules also can be detected by monitoring fluorescence energy transfer (FET) (see, for example, Lakowicz et al., U.S. Patent No. 5,631,169; Stavrianopoulos et al. U.S. Patent No. 4,868,103). A fluorophore label on a first, "donor" molecule is selected such that its emitted fluorescent energy will be absorbed by a fluorescent label on a second, "acceptor" molecule, which in turn is able to fluoresce due to the absorbed energy. Alternately, the "donor" polypeptide molecule may simply utilize the natural fluorescent energy of tryptophan residues. Labels are chosen that emit different wavelengths of light, such that the "acceptor" molecule label may be differentiated from that of the "donor". Since the efficiency of energy transfer between the labels is related to the distance separating the molecules, the spatial relationship between the molecules can be assessed. In a situation in which binding occurs between the molecules, the fluorescent emission of the "acceptor" molecule label in the assay should be maximal. An FET binding event can be conveniently measured through standard fluorometric detection means well known in the art (e.g., using a fluorimeter).

[0173] In another embodiment, determining the presence or absence of an interaction between a test molecule and a *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule can be effected by using real-time Biomolecular Interaction Analysis (BIA) (see, e.g., Sjolander & Urbaniczk, Anal. Chem. 63: 2338-2345 (1991) and Szabo et al., Curr. Opin. Struct. Biol. 5: 699-705 (1995)). "Surface plasmon resonance" or "BIA" detects biospecific interactions in real time, without labeling any of the interactants (e.g., BIAcore). Changes in the mass at the binding surface (indicative of a binding event) result in alterations of the refractive index of light near the surface (the optical phenomenon of surface plasmon resonance (SPR)), resulting in a detectable signal which can be used as an indication of real-time reactions between biological molecules.

[0174] In another embodiment, the *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule or test molecules are anchored to a solid phase. The *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule/test molecule complexes anchored to the solid phase can be detected at the end of the reaction. The target *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule is often anchored to a solid surface, and the test molecule, which is not anchored, can be labeled, either directly or indirectly, with detectable labels discussed herein.

[0175] It may be desirable to immobilize a *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule, an anti-*GP6, LAMA4, CHGB, LOC338749* or *TTN* antibody, or test molecules to facilitate separation of complexed from uncomplexed forms of *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecules and test molecules, as well as to accommodate automation of the assay. Binding of a test molecule to a *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule can be accomplished in any vessel suitable for containing the reactants. Examples of such vessels include microtiter plates, test tubes, and micro-centrifuge tubes. In one embodiment, a fusion polypeptide can be provided which adds a domain that allows a *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule to be bound to a matrix. For example, glutathione-S-transferase/*GP6, LAMA4, CHGB, LOC338749* or *TTN* fusion polypeptides or glutathione-S-transferase/target fusion polypeptides can be adsorbed onto glutathione sepharose beads (Sigma Chemical, St. Louis, MO) or glutathione derivitized microtiter plates, which are then combined with the test compound or the test compound and either the non-adsorbed target polypeptide or *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide, and the mixture incubated under conditions conducive to complex formation (e.g., at physiological conditions for salt and pH). Following incubation, the beads or microtiter plate wells are washed to remove any unbound components, the matrix immobilized in the case of beads, complex determined either directly or indirectly, for example, as described above. Alternatively, the complexes can be dissociated from the matrix, and the level of *GP6, LAMA4, CHGB, LOC338749* or *TTN* binding or activity determined using standard techniques.

[0176] Other techniques for immobilizing a *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule on matrices include using biotin and streptavidin. For example, biotinylated *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or target molecules can be prepared from biotin-NHS (N-hydroxy-succinimide) using techniques known in the art (e.g., biotinylation kit, Pierce Chemicals, Rockford, IL), and immobilized in the wells of streptavidin-coated 96 well plates (Pierce Chemical).

[0177] In order to conduct the assay, the non-immobilized component is added to the coated surface containing the anchored component. After the reaction is complete, unreacted components are removed (e.g., by washing) under conditions such that any complexes formed will remain immobilized on the solid surface. The detection of complexes anchored on the solid surface can be accomplished in a number of ways. Where the previously non-immobilized component is pre-labeled, the detection of label

immobilized on the surface indicates that complexes were formed. Where the previously non-immobilized component is not pre-labeled, an indirect label can be used to detect complexes anchored on the surface; e.g., using a labeled antibody specific for the immobilized component (the antibody, in turn, can be directly labeled or indirectly labeled with, e.g., a labeled anti-Ig antibody).

[0178] In one embodiment, this assay is performed utilizing antibodies reactive with *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or test molecules but which do not interfere with binding of the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide to its test molecule. Such antibodies can be derivitized to the wells of the plate, and unbound target or *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide trapped in the wells by antibody conjugation. Methods for detecting such complexes, in addition to those described above for the GST-immobilized complexes, include immunodetection of complexes using antibodies reactive with the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or target molecule, as well as enzyme-linked assays which rely on detecting an enzymatic activity associated with the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or test molecule.

[0179] Alternatively, cell free assays can be conducted in a liquid phase. In such an assay, the reaction products are separated from unreacted components, by any of a number of standard techniques, including but not limited to: differential centrifugation (see, for example, Rivas, G., and Minton, A. P., Trends Biochem Sci Aug;18(8): 284-7 (1993)); chromatography (gel filtration chromatography, ion-exchange chromatography); electrophoresis (see, e.g., Ausubel et al., eds. Current Protocols in Molecular Biology , J. Wiley: New York (1999)); and immunoprecipitation (see, for example, Ausubel, F. et al., eds. Current Protocols in Molecular Biology , J. Wiley: New York (1999)). Such resins and chromatographic techniques are known to one skilled in the art (see, e.g., Heegaard, J Mol. Recognit. Winter; 11(1-6): 141-8 (1998); Hage & Tweed, J. Chromatogr. B Biomed. Sci. Appl. Oct 10; 699 (1-2): 499-525 (1997)). Further, fluorescence energy transfer may also be conveniently utilized, as described herein, to detect binding without further purification of the complex from solution.

[0180] In another embodiment, modulators of *GP6, LAMA4, CHGB, LOC338749* or *TTN* expression are identified. For example, a cell or cell free mixture is contacted with a candidate compound and the expression of *GP6, LAMA4, CHGB, LOC338749* or *TTN* mRNA or polypeptide evaluated relative to the level of expression of *GP6, LAMA4, CHGB, LOC338749* or *TTN* mRNA or polypeptide in the absence of the candidate compound. When expression of *GP6, LAMA4, CHGB, LOC338749* or *TTN* mRNA or polypeptide is greater in the presence of the candidate compound than in its absence, the candidate compound is identified as a stimulator of *GP6, LAMA4, CHGB, LOC338749* or *TTN* mRNA or polypeptide expression. Alternatively, when expression of *GP6, LAMA4, CHGB, LOC338749* or *TTN* mRNA or polypeptide is less (statistically significantly less) in the presence of the candidate compound than in its absence, the candidate compound is identified as an inhibitor of *GP6, LAMA4, CHGB,*

LOC338749 or *TTN* mRNA or polypeptide expression. The level of *GP6, LAMA4, CHGB, LOC338749* or *TTN* mRNA or polypeptide expression can be determined by methods described herein for detecting *GP6, LAMA4, CHGB, LOC338749* or *TTN* mRNA or polypeptide.

[0181] In another embodiment, binding partners that interact with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule are detected. The *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecules can interact with one or more cellular or extracellular macromolecules, such as polypeptides, in vivo, and these molecules that interact with *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecules are referred to herein as "binding partners." Molecules that disrupt such interactions can be useful in regulating the activity of the target gene product. Such molecules can include, but are not limited to molecules such as antibodies, peptides, and small molecules. Target genes/products for use in this embodiment often are the *GP6, LAMA4, CHGB, LOC338749* or *TTN* genes herein identified. In an alternative embodiment, provided is a method for determining the ability of the test compound to modulate the activity of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide through modulation of the activity of a downstream effector of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* target molecule. For example, the activity of the effector molecule on an appropriate target can be determined, or the binding of the effector to an appropriate target can be determined, as previously described.

[0182] To identify compounds that interfere with the interaction between the target gene product and its cellular or extracellular binding partner(s), e.g., a substrate, a reaction mixture containing the target gene product and the binding partner is prepared, under conditions and for a time sufficient, to allow the two products to form complex. In order to test an inhibitory agent, the reaction mixture is provided in the presence and absence of the test compound. The test compound can be initially included in the reaction mixture, or can be added at a time subsequent to the addition of the target gene and its cellular or extracellular binding partner. Control reaction mixtures are incubated without the test compound or with a placebo. The formation of any complexes between the target gene product and the cellular or extracellular binding partner is then detected. The formation of a complex in the control reaction, but not in the reaction mixture containing the test compound, indicates that the compound interferes with the interaction of the target gene product and the interactive binding partner. Additionally, complex formation within reaction mixtures containing the test compound and normal target gene product can also be compared to complex formation within reaction mixtures containing the test compound and mutant target gene product. This comparison can be important in those cases where it is desirable to identify compounds that disrupt interactions of mutant but not normal target gene products.

[0183] These assays can be conducted in a heterogeneous or homogeneous format. Heterogeneous assays involve anchoring either the target gene product or the binding partner onto a solid phase, and detecting complexes anchored on the solid phase at the end of the reaction. In homogeneous assays, the

entire reaction is carried out in a liquid phase. In either approach, the order of addition of reactants can be varied to obtain different information about the compounds being tested. For example, test compounds that interfere with the interaction between the target gene products and the binding partners, e.g., by competition, can be identified by conducting the reaction in the presence of the test substance. Alternatively, test compounds that disrupt preformed complexes, e.g., compounds with higher binding constants that displace one of the components from the complex, can be tested by adding the test compound to the reaction mixture after complexes have been formed. The various formats are briefly described below.

[0184] In a heterogeneous assay system, either the target gene product or the interactive cellular or extracellular binding partner, is anchored onto a solid surface (e.g., a microtiter plate), while the non-anchored species is labeled, either directly or indirectly. The anchored species can be immobilized by non-covalent or covalent attachments. Alternatively, an immobilized antibody specific for the species to be anchored can be used to anchor the species to the solid surface.

[0185] In order to conduct the assay, the partner of the immobilized species is exposed to the coated surface with or without the test compound. After the reaction is complete, unreacted components are removed (e.g., by washing) and any complexes formed will remain immobilized on the solid surface. Where the non-immobilized species is pre-labeled, the detection of label immobilized on the surface indicates that complexes were formed. Where the non-immobilized species is not pre-labeled, an indirect label can be used to detect complexes anchored on the surface; e.g., using a labeled antibody specific for the initially non-immobilized species (the antibody, in turn, can be directly labeled or indirectly labeled with, e.g., a labeled anti-Ig antibody). Depending upon the order of addition of reaction components, test compounds that inhibit complex formation or that disrupt preformed complexes can be detected.

[0186] Alternatively, the reaction can be conducted in a liquid phase in the presence or absence of the test compound, the reaction products separated from unreacted components, and complexes detected; e.g., using an immobilized antibody specific for one of the binding components to anchor any complexes formed in solution, and a labeled antibody specific for the other partner to detect anchored complexes. Again, depending upon the order of addition of reactants to the liquid phase, test compounds that inhibit complex or that disrupt preformed complexes can be identified.

[0187] In an alternate embodiment, a homogeneous assay can be used. For example, a preformed complex of the target gene product and the interactive cellular or extracellular binding partner product is prepared in that either the target gene products or their binding partners are labeled, but the signal generated by the label is quenched due to complex formation (see, e.g., U.S. Patent No. 4,109,496 that utilizes this approach for immunoassays). The addition of a test substance that competes with and displaces one of the species from the preformed complex will result in the generation of a signal above

background. In this way, test substances that disrupt target gene product-binding partner interaction can be identified.

[0188] Also, binding partners of *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecules can be identified in a two-hybrid assay or three-hybrid assay (see, e.g., U.S. Patent No. 5,283,317; Zervos et al., Cell 72:223-232 (1993); Madura et al., J. Biol. Chem. 268: 12046-12054 (1993); Bartel et al., Biotechniques 14: 920-924 (1993); Iwabuchi et al., Oncogene 8: 1693-1696 (1993); and Brent WO94/10300), to identify other polypeptides, which bind to or interact with *GP6, LAMA4, CHGB, LOC338749* or *TTN* ("*GP6, LAMA4, CHGB, LOC338749* or *TTN*-binding polypeptides" or "*GP6, LAMA4, CHGB, LOC338749* or *TTN*-bp") and are involved in *GP6, LAMA4, CHGB, LOC338749* or *TTN* activity. Such *GP6, LAMA4, CHGB, LOC338749* or *TTN*-bps can be activators or inhibitors of signals by the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptides or *GP6, LAMA4, CHGB, LOC338749* or *TTN* targets as, for example, downstream elements of a *GP6, LAMA4, CHGB, LOC338749* or *TTN*-mediated signaling pathway.

[0189] A two-hybrid system is based on the modular nature of most transcription factors, which consist of separable DNA-binding and activation domains. Briefly, the assay utilizes two different DNA constructs. In one construct, the gene that codes for a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide is fused to a gene encoding the DNA binding domain of a known transcription factor (e.g., GAL-4). In the other construct, a DNA sequence, from a library of DNA sequences, that encodes an unidentified polypeptide ("prey" or "sample") is fused to a gene that codes for the activation domain of the known transcription factor. (Alternatively the: *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide can be the fused to the activator domain.) If the "bait" and the "prey" polypeptides are able to interact, in vivo, forming a *GP6, LAMA4, CHGB, LOC338749* or *TTN*-dependent complex, the DNA-binding and activation domains of the transcription factor are brought into close proximity. This proximity allows transcription of a reporter gene (e.g., LacZ) which is operably linked to a transcriptional regulatory site responsive to the transcription factor. Expression of the reporter gene can be detected and cell colonies containing the functional transcription factor can be isolated and used to obtain the cloned gene which encodes the polypeptide which interacts with the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide.

[0190] Candidate therapeutics for treating breast cancer are identified from a group of test molecules that interact with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid or polypeptide. Test molecules are normally ranked according to the degree with which they interact or modulate (e.g., agonize or antagonize) DNA replication and/or processing, RNA transcription and/or processing, polypeptide production and/or processing, and/or function of *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecules, for example, and then top ranking modulators are selected. In a preferred embodiment, the

candidate therapeutic (i.e., test molecule) acts as a *GP6, LAMA4, CHGB, LOC338749* or *TTN* antagonist. Also, pharmacogenomic information described herein can determine the rank of a modulator. Candidate therapeutics typically are formulated for administration to a subject.

Therapeutic Treatments

[0191] Formulations or pharmaceutical compositions typically include in combination with a pharmaceutically acceptable carrier, a compound, an antisense nucleic acid, a ribozyme, an antibody, a binding partner that interacts with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide, a *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid, or a fragment thereof. The formulated molecule may be one that is identified by a screening method described above. Also, formulations may comprise a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or fragment thereof. As used herein, the term "pharmaceutically acceptable carrier" includes solvents, dispersion media, coatings, antibacterial and antifungal agents, isotonic and absorption delaying agents, and the like, compatible with pharmaceutical administration. Supplementary active compounds can also be incorporated into the compositions.

[0192] A pharmaceutical composition is formulated to be compatible with its intended route of administration. Examples of routes of administration include parenteral, e.g., intravenous, intradermal, subcutaneous, oral (e.g., inhalation), transdermal (topical), transmucosal, and rectal administration. Solutions or suspensions used for parenteral, intradermal, or subcutaneous application can include the following components: a sterile diluent such as water for injection, saline solution, fixed oils, polyethylene glycols, glycerin, propylene glycol or other synthetic solvents; antibacterial agents such as benzyl alcohol or methyl parabens; antioxidants such as ascorbic acid or sodium bisulfite; chelating agents such as ethylenediaminetetraacetic acid; buffers such as acetates, citrates or phosphates and agents for the adjustment of tonicity such as sodium chloride or dextrose. pH can be adjusted with acids or bases, such as hydrochloric acid or sodium hydroxide. The parenteral preparation can be enclosed in ampoules, disposable syringes or multiple dose vials made of glass or plastic.

[0193] Oral compositions generally include an inert diluent or an edible carrier. For the purpose of oral therapeutic administration, the active compound can be incorporated with excipients and used in the form of tablets, troches, or capsules, e.g., gelatin capsules. Oral compositions can also be prepared using a fluid carrier for use as a mouthwash. Pharmaceutically compatible binding agents, and/or adjuvant materials can be included as part of the composition. The tablets, pills, capsules, troches and the like can contain any of the following ingredients, or compounds of a similar nature: a binder such as microcrystalline cellulose, gum tragacanth or gelatin; an excipient such as starch or lactose, a disintegrating agent such as alginic acid, Primogel, or corn starch; a lubricant such as magnesium stearate

or Sterotes; a glidant such as colloidal silicon dioxide; a sweetening agent such as sucrose or saccharin; or a flavoring agent such as peppermint, methyl salicylate, or orange flavoring.

[0194] Pharmaceutical compositions suitable for injectable use include sterile aqueous solutions (where water soluble) or dispersions and sterile powders for the extemporaneous preparation of sterile injectable solutions or dispersion. For intravenous administration, suitable carriers include physiological saline, bacteriostatic water, Cremophor EL™ (BASF, Parsippany, NJ) or phosphate buffered saline (PBS). In all cases, the composition must be sterile and should be fluid to the extent that easy syringability exists. It should be stable under the conditions of manufacture and storage and must be preserved against the contaminating action of microorganisms such as bacteria and fungi. The carrier can be a solvent or dispersion medium containing, for example, water, ethanol, polyol (for example, glycerol, propylene glycol, and liquid polyethylene glycol, and the like), and suitable mixtures thereof. The proper fluidity can be maintained, for example, by the use of a coating such as lecithin, by the maintenance of the required particle size in the case of dispersion and by the use of surfactants. Prevention of the action of microorganisms can be achieved by various antibacterial and antifungal agents, for example, parabens, chlorobutanol, phenol, ascorbic acid, thimerosal, and the like. In many cases, isotonic agents, for example, sugars, polyalcohols such as mannitol, sorbitol, sodium chloride sometimes are included in the composition. Prolonged absorption of the injectable compositions can be brought about by including in the composition an agent which delays absorption, for example, aluminum monostearate and gelatin.

[0195] Sterile injectable solutions can be prepared by incorporating the active compound in the required amount in an appropriate solvent with one or a combination of ingredients enumerated above, as required, followed by filtered sterilization. Generally, dispersions are prepared by incorporating the active compound into a sterile vehicle which contains a basic dispersion medium and the required other ingredients from those enumerated above. In the case of sterile powders for the preparation of sterile injectable solutions, methods of preparation often utilized are vacuum drying and freeze-drying which yields a powder of the active ingredient plus any additional desired ingredient from a previously sterile-filtered solution thereof.

[0196] For administration by inhalation, the compounds are delivered in the form of an aerosol spray from pressured container or dispenser which contains a suitable propellant, e.g., a gas such as carbon dioxide, or a nebulizer.

[0197] Systemic administration can also be by transmucosal or transdermal means. For transmucosal or transdermal administration, penetrants appropriate to the barrier to be permeated are used in the formulation. Such penetrants are generally known in the art, and include, for example, for transmucosal administration, detergents, bile salts, and fusidic acid derivatives. Transmucosal administration can be accomplished through the use of nasal sprays or suppositories. For transdermal

administration, the active compounds are formulated into ointments, salves, gels, or creams as generally known in the art. Molecules can also be prepared in the form of suppositories (e.g., with conventional suppository bases such as cocoa butter and other glycerides) or retention enemas for rectal delivery.

[0198] In one embodiment, active molecules are prepared with carriers that will protect the compound against rapid elimination from the body, such as a controlled release formulation, including implants and microencapsulated delivery systems. Biodegradable, biocompatible polymers can be used, such as ethylene vinyl acetate, polyanhydrides, polyglycolic acid, collagen, polyorthoesters, and polylactic acid. Methods for preparation of such formulations will be apparent to those skilled in the art. Materials can also be obtained commercially from Alza Corporation and Nova Pharmaceuticals, Inc. Liposomal suspensions (including liposomes targeted to infected cells with monoclonal antibodies to viral antigens) can also be used as pharmaceutically acceptable carriers. These can be prepared according to methods known to those skilled in the art, for example, as described in U.S. Patent No. 4,522,811.

[0199] It is advantageous to formulate oral or parenteral compositions in dosage unit form for ease of administration and uniformity of dosage. Dosage unit form as used herein refers to physically discrete units suited as unitary dosages for the subject to be treated; each unit containing a predetermined quantity of active compound calculated to produce the desired therapeutic effect in association with the required pharmaceutical carrier.

[0200] Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, e.g., for determining the LD_{50} (the dose lethal to 50% of the population) and the ED_{50} (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD_{50}/ED_{50}. Molecules which exhibit high therapeutic indices often are utilized. While molecules that exhibit toxic side effects may be used, care should be taken to design a delivery system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

[0201] The data obtained from the cell culture assays and animal studies can be used in formulating a range of dosage for use in humans. The dosage of such molecules often lies within a range of circulating concentrations that include the ED_{50} with little or no toxicity. The dosage may vary within this range depending upon the dosage form employed and the route of administration utilized. For any molecules used in the methods described herein, the therapeutically effective dose can be estimated initially from cell culture assays. A dose may be formulated in animal models to achieve a circulating plasma concentration range that includes the IC_{50} (i.e., the concentration of the test compound which achieves a half-maximal inhibition of symptoms) as determined in cell culture. Such information can be

used to more accurately determine useful doses in humans. Levels in plasma may be measured, for example, by high performance liquid chromatography.

[0202] As defined herein, a therapeutically effective amount of protein or polypeptide (i.e., an effective dosage) ranges from about 0.001 to 30 mg/kg body weight, sometimes about 0.01 to 25 mg/kg body weight, often about 0.1 to 20 mg/kg body weight, and more often about 1 to 10 mg/kg, 2 to 9 mg/kg, 3 to 8 mg/kg, 4 to 7 mg/kg, or 5 to 6 mg/kg body weight. The protein or polypeptide can be administered one time per week for between about 1 to 10 weeks, sometimes between 2 to 8 weeks, often between about 3 to 7 weeks, and more often for about 4, 5, or 6 weeks. The skilled artisan will appreciate that certain factors may influence the dosage and timing required to effectively treat a subject, including but not limited to the severity of the disease or disorder, previous treatments, the general health and/or age of the subject, and other diseases present. Moreover, treatment of a subject with a therapeutically effective amount of a protein, polypeptide, or antibody can include a single treatment, or sometimes can include a series of treatments.

[0203] With regard to polypeptide formulations, featured herein is a method for treating breast cancer in a subject, which comprises contacting one or more cells in the subject with a first *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide, where the subject comprises a second *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide having one or more polymorphic variations associated with cancer, and where the first polypeptide comprises fewer polymorphic variations associated with cancer than the second polypeptide. The first and second polypeptides are encoded by a nucleic acid which comprises a nucleotide sequence selected from the group consisting of the nucleotide sequence of SEQ ID NO: 1-11; a nucleotide sequence which encodes a polypeptide consisting of an amino acid sequence encoded by a nucleotide sequence of SEQ ID NO: 1-11; a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence of SEQ ID NO: 1-11 and a nucleotide sequence 90% or more identical to a nucleotide sequence of SEQ ID NO: 1-11. The subject is often a human.

[0204] For antibodies, a dosage of 0.1 mg/kg of body weight (generally 10 mg/kg to 20 mg/kg) is often utilized. If the antibody is to act in the brain, a dosage of 50 mg/kg to 100 mg/kg is often appropriate. Generally, partially human antibodies and fully human antibodies have a longer half-life within the human body than other antibodies. Accordingly, lower dosages and less frequent administration is often possible. Modifications such as lipidation can be used to stabilize antibodies and to enhance uptake and tissue penetration (e.g., into the brain). A method for lipidation of antibodies is described by Cruikshank et al., J. Acquired Immune Deficiency Syndromes and Human Retrovirology 14:193 (1997).

[0205] Antibody conjugates can be used for modifying a given biological response, the drug moiety is not to be construed as limited to classical chemical therapeutic agents. For example, the drug moiety may be a protein or polypeptide possessing a desired biological activity. Such proteins may include, for example, a toxin such as abrin, ricin A, pseudomonas exotoxin, or diphtheria toxin; a polypeptide such as tumor necrosis factor, .alpha.-interferon, .beta.-interferon, nerve growth factor, platelet derived growth factor, tissue plasminogen activator; or, biological response modifiers such as, for example, lymphokines, interleukin-1 ("IL-1"), interleukin-2 ("IL-2"), interleukin-6 ("IL-6"), granulocyte macrophage colony stimulating factor ("GM-CSF"), granulocyte colony stimulating factor ("G-CSF"), or other growth factors. Alternatively, an antibody can be conjugated to a second antibody to form an antibody heteroconjugate as described by Segal in U.S. Patent No. 4,676,980.

[0206] For compounds, exemplary doses include milligram or microgram amounts of the compound per kilogram of subject or sample weight, for example, about 1 microgram per kilogram to about 500 milligrams per kilogram, about 100 micrograms per kilogram to about 5 milligrams per kilogram, or about 1 microgram per kilogram to about 50 micrograms per kilogram. It is understood that appropriate doses of a small molecule depend upon the potency of the small molecule with respect to the expression or activity to be modulated. When one or more of these small molecules is to be administered to an animal (e.g., a human) in order to modulate expression or activity of a polypeptide or nucleic acid described herein, a physician, veterinarian, or researcher may, for example, prescribe a relatively low dose at first, subsequently increasing the dose until an appropriate response is obtained. In addition, it is understood that the specific dose level for any particular animal subject will depend upon a variety of factors including the activity of the specific compound employed, the age, body weight, general health, gender, and diet of the subject, the time of administration, the route of administration, the rate of excretion, any drug combination, and the degree of expression or activity to be modulated.

[0207] *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid molecules can be inserted into vectors and used in gene therapy methods for treating breast cancer. Featured herein is a method for treating breast cancer in a subject, which comprises contacting one or more cells in the subject with a first *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid, where genomic DNA in the subject comprises a second *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid comprising one or more polymorphic variations associated with breast cancer, and where the first *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid comprises fewer polymorphic variations associated with breast cancer. The first and second nucleic acids typically comprise a nucleotide sequence selected from the group consisting of the nucleotide sequence of SEQ ID NO: 1-11; a nucleotide sequence which encodes a polypeptide consisting of an amino acid sequence encoded by a nucleotide sequence of SEQ ID NO: 1-11; a nucleotide sequence that is 90% or more identical to the nucleotide sequence of SEQ ID NO: 1-11, and a nucleotide sequence

which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence of SEQ ID NO: 1-11. The subject often is a human.

[0208] Gene therapy vectors can be delivered to a subject by, for example, intravenous injection, local administration (see U.S. Patent 5,328,470) or by stereotactic injection (see e.g., Chen et al., (1994) Proc. Natl. Acad. Sci. USA 91:3054-3057). Pharmaceutical preparations of gene therapy vectors can include a gene therapy vector in an acceptable diluent, or can comprise a slow release matrix in which the gene delivery vehicle is imbedded. Alternatively, where the complete gene delivery vector can be produced intact from recombinant cells (e.g., retroviral vectors) the pharmaceutical preparation can include one or more cells which produce the gene delivery system. Examples of gene delivery vectors are described herein.

[0209] Pharmaceutical compositions can be included in a container, pack, or dispenser together with instructions for administration.

[0210] Pharmaceutical compositions of active ingredients can be administered by any of the paths described herein for therapeutic and prophylactic methods for treating breast cancer. With regard to both prophylactic and therapeutic methods of treatment, such treatments may be specifically tailored or modified, based on knowledge obtained from pharmacogenomic analyses described herein. As used herein, the term "treatment" is defined as the application or administration of a therapeutic agent to a patient, or application or administration of a therapeutic agent to an isolated tissue or cell line from a patient, who has a disease, a symptom of disease or a predisposition toward a disease, with the purpose to cure, heal, alleviate, relieve, alter, remedy, ameliorate, improve or affect the disease, the symptoms of disease or the predisposition toward disease. A therapeutic agent includes, but is not limited to, small molecules, peptides, antibodies, ribozymes and antisense oligonucleotides.

[0211] Administration of a prophylactic agent can occur prior to the manifestation of symptoms characteristic of the *GP6, LAMA4, CHGB, LOC338749* or *TTN* aberrance, such that a disease or disorder is prevented or, alternatively, delayed in its progression. Depending on the type of *GP6, LAMA4, CHGB, LOC338749* or *TTN* aberrance, for example, a *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecule, *GP6, LAMA4, CHGB, LOC338749* or *TTN* agonist, or *GP6, LAMA4, CHGB, LOC338749* or *TTN* antagonist agent can be used for treating the subject. The appropriate agent can be determined based on screening assays described herein.

[0212] As discussed, successful treatment of *GP6, LAMA4, CHGB, LOC338749* or *TTN* disorders can be brought about by techniques that serve to inhibit the expression or activity of target gene products. For example, compounds (e.g., an agent identified using an assays described above) that exhibit negative modulatory activity can be used to prevent and/or treat breast cancer. Such molecules can include, but are not limited to peptides, phosphopeptides, small organic or inorganic molecules, or antibodies

(including, for example, polyclonal, monoclonal, humanized, anti-idiotypic, chimeric or single chain antibodies, and FAb, F(ab')2 and FAb expression library fragments, scFV molecules, and epitope-binding fragments thereof).

[0213] Further, antisense and ribozyme molecules that inhibit expression of the target gene can also be used to reduce the level of target gene expression, thus effectively reducing the level of target gene activity. Still further, triple helix molecules can be utilized in reducing the level of target gene activity. Antisense, ribozyme and triple helix molecules are discussed above.

[0214] It is possible that the use of antisense, ribozyme, and/or triple helix molecules to reduce or inhibit mutant gene expression can also reduce or inhibit the transcription (triple helix) and/or translation (antisense, ribozyme) of mRNA produced by normal target gene alleles, such that the concentration of normal target gene product present can be lower than is necessary for a normal phenotype. In such cases, nucleic acid molecules that encode and express target gene polypeptides exhibiting normal target gene activity can be introduced into cells via gene therapy method. Alternatively, in instances where the target gene encodes an extracellular polypeptide, normal target gene polypeptide often is co-administered into the cell or tissue to maintain the requisite level of cellular or tissue target gene activity.

[0215] Another method by which nucleic acid molecules may be utilized in treating or preventing a disease characterized by *GP6, LAMA4, CHGB, LOC338749* or *TTN* expression is through the use of aptamer molecules specific for *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. Aptamers are nucleic acid molecules having a tertiary structure which permits them to specifically bind to polypeptide ligands (see, e.g., Osborne, et al., Curr. Opin. Chem. Biol.1(1): 5-9 (1997); and Patel, D. J., Curr. Opin. Chem. Biol. Jun;1(1): 32-46 (1997)). Since nucleic acid molecules may in many cases be more conveniently introduced into target cells than therapeutic polypeptide molecules may be, aptamers offer a method by which *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide activity may be specifically decreased without the introduction of drugs or other molecules which may have pluripotent effects.

[0216] Antibodies can be generated that are both specific for target gene product and that reduce target gene product activity. Such antibodies may, therefore, by administered in instances whereby negative modulatory techniques are appropriate for the treatment of *GP6, LAMA4, CHGB, LOC338749* or *TTN* disorders. For a description of antibodies, see the Antibody section above.

[0217] In circumstances where injection of an animal or a human subject with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or epitope for stimulating antibody production is harmful to the subject, it is possible to generate an immune response against *GP6, LAMA4, CHGB, LOC338749* or *TTN* through the use of anti-idiotypic antibodies (see, for example, Herlyn, D., Ann. Med.;31(1): 66-78 (1999); and Bhattacharya-Chatterjee & Foon, Cancer Treat. Res.; 94: 51-68 (1998)). If an anti-idiotypic antibody is introduced into a mammal or human subject, it should stimulate the production of anti-anti-

idiotypic antibodies, which should be specific to the *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide. Vaccines directed to a disease characterized by *GP6, LAMA4, CHGB, LOC338749* or *TTN* expression may also be generated in this fashion.

[0218] In instances where the target antigen is intracellular and whole antibodies are used, internalizing antibodies may be utilized. Lipofectin or liposomes can be used to deliver the antibody or a fragment of the Fab region that binds to the target antigen into cells. Where fragments of the antibody are used, the smallest inhibitory fragment that binds to the target antigen often is utilized. For example, peptides having an amino acid sequence corresponding to the Fv region of the antibody can be used. Alternatively, single chain neutralizing antibodies that bind to intracellular target antigens can also be administered. Such single chain antibodies can be administered, for example, by expressing nucleotide sequences encoding single-chain antibodies within the target cell population (see e.g., Marasco et al., Proc. Natl. Acad. Sci. USA 90: 7889-7893 (1993)).

[0219] *GP6, LAMA4, CHGB, LOC338749* or *TTN* molecules and compounds that inhibit target gene expression, synthesis and/or activity can be administered to a patient at therapeutically effective doses to prevent, treat or ameliorate *GP6, LAMA4, CHGB, LOC338749* or *TTN* disorders. A therapeutically effective dose refers to that amount of the compound sufficient to result in amelioration of symptoms of the disorders.

[0220] Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, e.g., for determining the LD_{50} (the dose lethal to 50% of the population) and the ED_{50} (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD_{50}/ED_{50}. Compounds that exhibit large therapeutic indices often are utilized. While compounds that exhibit toxic side effects can be used, care should be taken to design a delivery system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

[0221] Data obtained from cell culture assays and animal studies can be used in formulating a range of dosage for use in humans. The dosage of such compounds often lies within a range of circulating concentrations that include the ED_{50} with little or no toxicity. The dosage can vary within this range depending upon the dosage form employed and the route of administration utilized. For any compound used in a method described herein, the therapeutically effective dose can be estimated initially from cell culture assays. A dose can be formulated in animal models to achieve a circulating plasma concentration range that includes the IC_{50} (i.e., the concentration of the test compound that achieves a half-maximal inhibition of symptoms) as determined in cell culture. Such information can be used to more accurately

determine useful doses in humans. Levels in plasma can be measured, for example, by high performance liquid chromatography.

[0222] Another example of effective dose determination for an individual is the ability to directly assay levels of "free" and "bound" compound in the serum of the test subject. Such assays may utilize antibody mimics and/or "biosensors" that have been created through molecular imprinting techniques. The compound which is able to modulate *GP6, LAMA4, CHGB, LOC338749* or *TTN* activity is used as a template, or "imprinting molecule", to spatially organize polymerizable monomers prior to their polymerization with catalytic reagents. The subsequent removal of the imprinted molecule leaves a polymer matrix which contains a repeated "negative image" of the compound and is able to selectively rebind the molecule under biological assay conditions. A detailed review of this technique can be seen in Ansell et al., Current Opinion in Biotechnology 7: 89-94 (1996) and in Shea, Trends in Polymer Science 2: 166-173 (1994). Such "imprinted" affinity matrixes are amenable to ligand-binding assays, whereby the immobilized monoclonal antibody component is replaced by an appropriately imprinted matrix. An example of the use of such matrixes in this way can be seen in Vlatakis, et al., Nature 361: 645-647 (1993). Through the use of isotope-labeling, the "free" concentration of compound which modulates the expression or activity of *GP6, LAMA4, CHGB, LOC338749* or *TTN* can be readily monitored and used in calculations of IC_{50}. Such "imprinted" affinity matrixes can also be designed to include fluorescent groups whose photon-emitting properties measurably change upon local and selective binding of target compound. These changes can be readily assayed in real time using appropriate fiberoptic devices, in turn allowing the dose in a test subject to be quickly optimized based on its individual IC_{50}. A rudimentary example of such a "biosensor" is discussed in Kriz et al., Analytical Chemistry 67: 2142-2144 (1995).

[0223] Provided herein are methods of modulating *GP6, LAMA4, CHGB, LOC338749* or *TTN* expression or activity for therapeutic purposes. Accordingly, in an exemplary embodiment, the modulatory method involves contacting a cell with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* or agent that modulates one or more of the activities of *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide activity associated with the cell. An agent that modulates *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide activity can be an agent as described herein, such as a nucleic acid or a polypeptide, a naturally-occurring target molecule of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide (e.g., a *GP6, LAMA4, CHGB, LOC338749* or *TTN* substrate or receptor), a *GP6, LAMA4, CHGB, LOC338749* or *TTN* antibody, a *GP6, LAMA4, CHGB, LOC338749* or *TTN* agonist or antagonist, a peptidomimetic of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* agonist or antagonist, or other small molecule.

[0224] In one embodiment, the agent stimulates one or more *GP6, LAMA4, CHGB, LOC338749* or *TTN* activities. Examples of such stimulatory agents include active *GP6, LAMA4, CHGB, LOC338749*

or *TTN* polypeptide and a nucleic acid molecule encoding *GP6, LAMA4, CHGB, LOC338749* or *TTN*. In another embodiment, the agent inhibits one or more *GP6, LAMA4, CHGB, LOC338749* or *TTN* activities. Examples of such inhibitory agents include antisense *GP6, LAMA4, CHGB, LOC338749* or *TTN* nucleic acid molecules, anti-*GP6, LAMA4, CHGB, LOC338749* or *TTN* antibodies, and *GP6, LAMA4, CHGB, LOC338749* or *TTN* inhibitors. These modulatory methods can be performed in vitro (e.g., by culturing the cell with the agent) or, alternatively, in vivo (e.g., by administering the agent to a subject). As such, provided are methods of treating an individual afflicted with a disease or disorder characterized by aberrant or unwanted expression or activity of a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or nucleic acid molecule. In one embodiment, the method involves administering an agent (e.g., an agent identified by a screening assay described herein), or combination of agents that modulates (e.g., upregulates or downregulates) *GP6, LAMA4, CHGB, LOC338749* or *TTN* expression or activity. In a preferred embodiment, the method involves administering an agent (e.g., an agent identified by a screening assay described herein), or combination of agents that inhibits *GP6, LAMA4, CHGB, LOC338749* or *TTN* expression or activity. In another embodiment, the method involves administering a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polypeptide or nucleic acid molecule as therapy to compensate for reduced, aberrant, or unwanted *GP6, LAMA4, CHGB, LOC338749* or *TTN* expression or activity.

[0225] Stimulation of *GP6, LAMA4, CHGB, LOC338749* or *TTN* activity is desirable in situations in which *GP6, LAMA4, CHGB, LOC338749* or *TTN* is abnormally downregulated and/or in which increased *GP6, LAMA4, CHGB, LOC338749* or *TTN* activity is likely to have a beneficial effect. For example, stimulation of *GP6, LAMA4, CHGB, LOC338749* or *TTN* activity is desirable in situations in which a *GP6, LAMA4, CHGB, LOC338749* or *TTN* is downregulated and/or in which increased *GP6, LAMA4, CHGB, LOC338749* or *TTN* activity is likely to have a beneficial effect. Likewise, inhibition of *GP6, LAMA4, CHGB, LOC338749* or *TTN* activity is desirable in situations in which *GP6, LAMA4, CHGB, LOC338749* or *TTN* is abnormally upregulated and/or in which decreased *GP6, LAMA4, CHGB, LOC338749* or *TTN* activity is likely to have a beneficial effect.

Methods of Treatment

[0226] In another aspect, provided are methods for identifying a risk of cancer in an individual as described herein and, if a genetic predisposition is identified, treating that individual to delay or reduce or prevent the development of cancer. Such a procedure can be used to treat breast cancer. Optionally, treating an individual for cancer may include inhibiting cellular proliferation, inhibiting metastasis, inhibiting invasion, or preventing tumor formation or growth as defined herein. Suitable treatments to prevent or reduce or delay breast cancer focus on inhibiting additional cellular proliferation, inhibiting

metastasis, inhibiting invasion, and preventing further tumor formation or growth. Treatment usually includes surgery followed by radiation therapy. Surgery may be a lumpectomy or a mastectomy (e.g., total, simple or radical). Even if the doctor removes all of the cancer that can be seen at the time of surgery, the patient may be given radiation therapy, chemotherapy, or hormone therapy after surgery to try to kill any cancer cells that may be left. Radiation therapy is the use of x-rays or other types of radiation to kill cancer cells and shrink tumors. Radiation therapy may use external radiation (using a machine outside the body) or internal radiation. Chemotherapy is the use of drugs to kill cancer cells. Chemotherapy may be taken by mouth, or it may be put into the body by inserting a needle into a vein or muscle. Hormone therapy often focuses on estrogen and progesterone, which are hormones that affect the way some cancers grow. If tests show that the cancer cells have estrogen and progesterone receptors (molecules found in some cancer cells to which estrogen and progesterone will attach), hormone therapy is used to block the way these hormones help the cancer grow. Hormone therapy with tamoxifen is often given to patients with early stages of breast cancer and those with metastatic breast cancer. Other types of treatment being tested in clinical trials include sentinel lymph node biopsy followed by surgery and high-dose chemotherapy with bone marrow transplantation and peripheral blood stem cell transplantation. Any preventative/therapeutic treatment known in the art may be prescribed and/or administered, including, for example, surgery, chemotherapy and/or radiation treatment, and any of the treatments may be used in combination with one another to treat or prevent breast cancer (e.g., surgery followed by radiation therapy).

[0227] Also provided are methods of preventing or treating cancer comprising providing an individual in need of such treatment with a *GP6, LAMA4, CHGB, LOC338749* or *TTN* inhibitor that reduces or inhibits the overexpression of mutant *GP6, LAMA4, CHGB, LOC338749* or *TTN* (e.g., a *GP6, LAMA4, CHGB, LOC338749* or *TTN* polynucleotide with an allele that is associated with cancer). Included herein are methods of reducing or blocking the expression of *GP6, LAMA4, CHGB, LOC338749* or *TTN* comprising providing or administering to individuals in need of reducing or blocking the expression of *GP6, LAMA4, CHGB, LOC338749* or *TTN* a pharmaceutical or physiologically acceptable composition comprising a molecule capable of inhibiting expression of *GP6, LAMA4, CHGB, LOC338749* or *TTN*, e.g., a siRNA molecule. Also included herein are methods of reducing or blocking the expression of secondary regulatory genes regulated by *GP6, LAMA4, CHGB, LOC338749* or *TTN* that play a role in oncogenesis which comprises introducing competitive inhibitors that target *GP6, LAMA4, CHGB, LOC338749* or *TTN*'s effect on these regulatory genes or that block the binding of positive factors necessary for the expression of these regulatory genes.

[0228] The examples set forth below are intended to illustrate but not limit the invention.

Examples

[0229] In the following studies a group of subjects were selected according to specific parameters relating to breast cancer. Nucleic acid samples obtained from individuals in the study group were subjected to genetic analysis, which identified associations between breast cancer and certain polymorphic regions in the GP6, LAMA4, CHGB/C20orf154, LOC338749 and TTN/LOC351327 genes or gene regions (herein referred to as "target genes", "target nucleotides", "target polypeptides" or simply "targets"). Methods are described for producing *GP6, LAMA4, CHGB/C20orf154, LOC338749*, or *TTN/LOC351327* polypeptides and polypeptide variants *in vitro* or *in vivo*. *GP6, LAMA4, CHGB/C20orf154, LOC338749* or *TTN/LOC351327* nucleic acids or polypeptides and variants thereof are utilized for screening test molecules for those that interact with *GP6, LAMA4, CHGB/C20orf154, LOC338749* or *TTN/LOC351327* molecules. Test molecules identified as interactors with *GP6, LAMA4, CHGB/C20orf154, LOC338749* or *TTN/LOC351327* molecules and variants are further screened *in vivo* to determine whether they treat breast cancer.

Example 1

Samples and Pooling Strategies

Sample Selection

[0230] Blood samples were collected from individuals diagnosed with breast cancer, which were referred to as case samples. Also, blood samples were collected from individuals not diagnosed with breast cancer as gender and age-matched controls. All of the samples were of German/German descent. A database was created that listed all phenotypic trait information gathered from individuals for each case and control sample. Genomic DNA was extracted from each of the blood samples for genetic analyses.

DNA Extraction from Blood Samples

[0231] Six to ten milliliters of whole blood was transferred to a 50 ml tube containing 27 ml of red cell lysis solution (RCL). The tube was inverted until the contents were mixed. Each tube was incubated for 10 minutes at room temperature and inverted once during the incubation. The tubes were then centrifuged for 20 minutes at 3000 x g and the supernatant was carefully poured off. 100-200 µl of residual liquid was left in the tube and was pipetted repeatedly to resuspend the pellet in the residual supernatant. White cell lysis solution (WCL) was added to the tube and pipetted repeatedly until completely mixed. While no incubation was normally required, the solution was incubated at 37°C or room temperature if cell clumps were visible after mixing until the solution was homogeneous. 2 ml of protein precipitation was added to the cell lysate. The mixtures were vortexed vigorously at high speed

for 20 sec to mix the protein precipitation solution uniformly with the cell lysate, and then centrifuged for 10 minutes at 3000 x g. The supernatant containing the DNA was then poured into a clean 15 ml tube, which contained 7 ml of 100% isopropanol. The samples were mixed by inverting the tubes gently until white threads of DNA were visible. Samples were centrifuged for 3 minutes at 2000 x g and the DNA was visible as a small white pellet. The supernatant was decanted and 5 ml of 70% ethanol was added to each tube. Each tube was inverted several times to wash the DNA pellet, and then centrifuged for 1 minute at 2000 x g. The ethanol was decanted and each tube was drained on clean absorbent paper. The DNA was dried in the tube by inversion for 10 minutes, and then 1000 μl of 1X TE was added. The size of each sample was estimated, and less TE buffer was added during the following DNA hydration step if the sample was smaller. The DNA was allowed to rehydrate overnight at room temperature, and DNA samples were stored at 2-8°C.

[0232] DNA was quantified by placing samples on a hematology mixer for at least 1 hour. DNA was serially diluted (typically 1:80, 1:160, 1:320, and 1:640 dilutions) so that it would be within the measurable range of standards. 125 μl of diluted DNA was transferred to a clear U-bottom microtitre plate, and 125 μl of 1X TE buffer was transferred into each well using a multichannel pipette. The DNA and 1X TE were mixed by repeated pipetting at least 15 times, and then the plates were sealed. 50 μl of diluted DNA was added to wells A5-H12 of a black flat bottom microtitre plate. Standards were inverted six times to mix them, and then 50 μl of 1X TE buffer was pipetted into well A1, 1000 ng/ml of standard was pipetted into well A2, 500 ng/ml of standard was pipetted into well A3, and 250 ng/ml of standard was pipetted into well A4. PicoGreen (Molecular Probes, Eugene, Oregon) was thawed and freshly diluted 1:200 according to the number of plates that were being measured. PicoGreen was vortexed and then 50μl was pipetted into all wells of the black plate with the diluted DNA. DNA and PicoGreen were mixed by pipetting repeatedly at least 10 times with the multichannel pipette. The plate was placed into a Fluoroskan Ascent Machine (microplate fluorometer produced by Labsystems) and the samples were allowed to incubate for 3 minutes before the machine was run using filter pairs 485 nm excitation and 538 nm emission wavelengths. Samples having measured DNA concentrations of greater than 450 ng/μl were re-measured for conformation. Samples having measured DNA concentrations of 20 ng/μl or less were re-measured for confirmation.

Pooling Strategies

[0233] Samples were placed into one of two groups based on disease status. The two groups were female case groups and female control groups. A select set of samples from each group were utilized to generate pools, and one pool was created for each group. Each individual sample in a pool was represented by an equal amount of genomic DNA. For example, where 25 ng of genomic DNA was

utilized in each PCR reaction and there were 200 individuals in each pool, each individual would provide 125 pg of genomic DNA. Inclusion or exclusion of samples for a pool was based upon the following criteria: the sample was derived from an individual characterized as Caucasian; the sample was derived from an individual of German paternal and maternal descent; the database included relevant phenotype information for the individual; case samples were derived from individuals diagnosed with breast cancer; control samples were derived from individuals free of cancer and no family history of breast cancer; and sufficient genomic DNA was extracted from each blood sample for all allelotyping and genotyping reactions performed during the study. Phenotype information included pre- or post-menopausal, familial predisposition, country or origin of mother and father, diagnosis with breast cancer (date of primary diagnosis, age of individual as of primary diagnosis, grade or stage of development, occurrence of metastases, e.g., lymph node metastases, organ metastases), condition of body tissue (skin tissue, breast tissue, ovary tissue, peritoneum tissue and myometrium), method of treatment (surgery, chemotherapy, hormone therapy, radiation therapy). Samples that met these criteria were added to appropriate pools based on gender and disease status.

[0234] The selection process yielded the pools set forth in Table 1, which were used in the studies that follow:

TABLE 1

	Female CASE	**Female CONTROL**
Pool size (Number)	272	276
Pool Criteria (ex: case/control)	case	control
Mean Age (ex: years)	59.6	55.4

Example 2

Association of Polymorphic Variants with Breast cancer

[0235] A whole-genome screen was performed to identify particular SNPs associated with occurrence of breast cancer. As described in Example 1, two sets of samples were utilized, which included samples from female individuals having breast cancer (breast cancer cases) and samples from female individuals not having cancer (female controls). The initial screen of each pool was performed in an allelotyping study, in which certain samples in each group were pooled. By pooling DNA from each group, an allele frequency for each SNP in each group was calculated. These allele frequencies were then compared to one another. Particular SNPs were considered as being associated with breast cancer

when allele frequency differences calculated between case and control pools were statistically significant. SNP disease association results obtained from the allelotyping study were then validated by genotyping each associated SNP across all samples from each pool. The results of the genotyping were then analyzed, allele frequencies for each group were calculated from the individual genotyping results, and a p-value was calculated to determine whether the case and control groups had statistically significantly differences in allele frequencies for a particular SNP. When the genotyping results agreed with the original allelotyping results, the SNP disease association was considered validated at the genetic level.

SNP Panel Used for Genetic Analyses

[0236] A whole-genome SNP screen began with an initial screen of approximately 25,000 SNPs over each set of disease and control samples using a pooling approach. The pools studied in the screen are described in Example 1. The SNPs analyzed in this study were part of a set of 25,488 SNPs confirmed as being statistically polymorphic as each is characterized as having a minor allele frequency of greater than 10%. The SNPs in the set reside in genes or in close proximity to genes, and many reside in gene exons. Specifically, SNPs in the set are located in exons, introns, and within 5,000 base-pairs upstream of a transcription start site of a gene. In addition, SNPs were selected according to the following criteria: they are located in ESTs; they are located in Locuslink or Ensemble genes; and they are located in Genomatix promoter predictions. SNPs in the set also were selected on the basis of even spacing across the genome, as depicted in Table 2.

[0237] A case-control study design using a whole genome association strategy involving approximately 28,000 single nucleotide polymorphisms (SNPs) was employed. Approximately 25,000 SNPs were evenly spaced in gene-based regions of the human genome with a median inter-marker distance of about 40,000 base pairs. Additionally, approximately 3,000 SNPs causing amino acid substitutions in genes described in the literature as candidates for various diseases were used. The case-control study samples were of female German origin (German paternal and maternal descent) 548 individuals were equally distributed in two groups (female controls and female cases). The whole genome association approach was first conducted on 2 DNA pools representing the 2 groups. Significant markers were confirmed by individual genotyping.

TABLE 2

General Statistics		Spacing Statistics	
Total # of SNPs	25,488	Median	37,058 bp
# of Exonic SNPs	>4,335 (17%)	Minimum*	1,000 bp
# SNPs with refSNP ID	20,776 (81%)	Maximum*	3,000,000 bp
Gene Coverage	>10,000	Mean	122,412 bp
Chromosome Coverage	All	Std Deviation	373,325 bp
		**Excludes outliers*	

Allelotyping and Genotyping Results

[0238] The genetic studies summarized above and described in more detail below identified allelic variants associated with breast cancer. The allelic variants identified from the SNP panel described in Table 2 are summarized below in Table 3.

Table 3

SNP Reference	Chromo-some Position	Position in Figs 1-4	Contig Identification	Contig Position	Sequence Identification	Sequence Position	Allelic Variability
rs1671152	60202366	45666	NT_011109	27794535	NM_016363	exonic (T323K)	T/G
rs1050348	112494002	47502	NT_025741	16663301	NM_002290	exonic (H491Y)	C/T
rs454422	5891693	49293	NT_011387	5883693	NM_032485	intragenic	A/C
rs763471	10491273	49273	NT_009237	1853223			G/T
rs2046778	179636570	49170	NT_005403	29831884	X90569	upstream	A/G

[0239] Table 3 includes information pertaining to the incident polymorphic variant associated with breast cancer identified herein. Public information pertaining to the polymorphism and the genomic sequence that includes the polymorphism are indicated. The genomic sequences identified in Table 3 may be accessed at the http address www.ncbi.nih.gov/entrez/query.fcgi, for example, by using the publicly available SNP reference number. The chromosome position refers to the position of the SNP within NCBI's Genome Build 33, which may be accessed at the following http address: www.ncbi.nlm.nih.gov/mapview/map_search.cgi?chr=hum_chr.inf&query=. The "Contig Position" provided in Table 3 corresponds to a nucleotide position set forth in the contig sequence, and designates the polymorphic site corresponding to the SNP reference number. The sequence containing the polymorphisms also may be referenced by the "Sequence Identification" set forth in Table 3. The "Sequence Identification" corresponds to cDNA sequence that encodes associated target polypeptides of

the invention. The position of the SNP within the cDNA sequence is provided in the "Sequence Position" column of Table 3. Also, the allelic variation at the polymorphic site and the allelic variant identified as associated with breast cancer is specified in Table 3. All nucleotide sequences referenced and accessed by the parameters set forth in Table 3 are incorporated herein by reference.

Assay for Verifying, Allelotyping, and Genotyping SNPs

[0240] A MassARRAY™ system (Sequenom, Inc.) was utilized to perform SNP genotyping in a high-throughput fashion. This genotyping platform was complemented by a homogeneous, single-tube assay method (hME™ or homogeneous MassEXTEND™ (Sequenom, Inc.)) in which two genotyping primers anneal to and amplify a genomic target surrounding a polymorphic site of interest. A third primer (the MassEXTEND™ primer), which is complementary to the amplified target up to but not including the polymorphism, was then enzymatically extended one or a few bases through the polymorphic site and then terminated.

[0241] For each polymorphism, SpectroDESIGNER™ software (Sequenom, Inc.) was used to generate a set of PCR primers and a MassEXTEND™ primer was used to genotype the polymorphism. Table 4 shows PCR primers and Table 5 shows extension primers used for analyzing polymorphisms. The initial PCR amplification reaction was performed in a 5 μl total volume containing 1X PCR buffer with 1.5 mM $MgCl_2$ (Qiagen), 200 μM each of dATP, dGTP, dCTP, dTTP (Gibco-BRL), 2.5 ng of genomic DNA, 0.1 units of HotStar DNA polymerase (Qiagen), and 200 nM each of forward and reverse PCR primers specific for the polymorphic region of interest.

TABLE 4: PCR Primers

Reference SNP ID	Forward PCR primer	Reverse PCR primer
rs1671152	ACGTTGGATGAGGGCTGTGCAGAGGCCGCTT	ACGTTGGATGTGAACATCCTGTCGGCCTCC
rs1050348	CAGCTGGATGACTACAATGC	TGTTCATGTCTTCGGCATCC
rs454422	CAGCTTTTGAGGCACTTTCC	AGCACCTTGCATACCCATAG
rs763471	TAACTCCTGTGTGGCTTTCT	GTGAAGAGCTCTGAAATGCC
rs2046778	CATGAAGCCTTATGCTTGAG	GTTCCCTTCCCCCATAAAAC

[0242] Samples were incubated at 95°C for 15 minutes, followed by 45 cycles of 95°C for 20 seconds, 56°C for 30 seconds, and 72°C for 1 minute, finishing with a 3 minute final extension at 72°C. Following amplification, shrimp alkaline phosphatase (SAP) (0.3 units in a 2 μl volume) (Amersham Pharmacia) was added to each reaction (total reaction volume was 7 μl) to remove any residual dNTPs

that were not consumed in the PCR step. Samples were incubated for 20 minutes at 37°C, followed by 5 minutes at 85°C to denature the SAP.

[0243] Once the SAP reaction was complete, a primer extension reaction was initiated by adding a polymorphism-specific MassEXTEND™ primer cocktail to each sample. Each MassEXTEND™ cocktail included a specific combination of dideoxynucleotides (ddNTPs) and deoxynucleotides (dNTPs) used to distinguish polymorphic alleles from one another. In Table 5, ddNTPs are shown and the fourth nucleotide not shown is the dNTP.

TABLE 5: Extend Primers

Reference SNP ID	Extend Probe	Term Mix
rs1671152	CTCCATCCTGACCCCCGT	ACT
rs1050348	CACTTGACCAGGCCCTTAAC	ACG
rs454422	GATCCTTCTCACTTACTGTTC	ACT
rs763471	CTCCAAGCAGTAAAGATGTTC	CGT
rs2046778	CTGTCATGATTGACAGGTCC	ACT

[0244] The MassEXTEND™ reaction was performed in a total volume of 9 µl, with the addition of 1X ThermoSequenase buffer, 0.576 units of ThermoSequenase (Amersham Pharmacia), 600 nM MassEXTEND™ primer, 2 mM of ddATP and/or ddCTP and/or ddGTP and/or ddTTP, and 2 mM of dATP or dCTP or dGTP or dTTP. The deoxy nucleotide (dNTP) used in the assay normally was complementary to the nucleotide at the polymorphic site in the amplicon. Samples were incubated at 94°C for 2 minutes, followed by 55 cycles of 5 seconds at 94°C, 5 seconds at 52°C, and 5 seconds at 72°C.

[0245] Following incubation, samples were desalted by adding 16 µl of water (total reaction volume was 25 µl), 3 mg of SpectroCLEAN™ sample cleaning beads (Sequenom, Inc.) and allowed to incubate for 3 minutes with rotation. Samples were then robotically dispensed using a piezoelectric dispensing device (SpectroJET™ (Sequenom, Inc.)) onto either 96-spot or 384-spot silicon chips containing a matrix that crystallized each sample (SpectroCHIP® (Sequenom, Inc.)). Subsequently, MALDI-TOF mass spectrometry (Biflex and Autoflex MALDI-TOF mass spectrometers (Bruker Daltonics) can be used) and SpectroTYPER RT™ software (Sequenom, Inc.) were used to analyze and interpret the SNP genotype for each sample.

Genetic Analysis

[0246] Variations identified in the target genes are provided in their respective genomic sequences (see Figures 1-5) Minor allelic frequencies for these polymorphisms was verified as being 10% or greater by determining the allelic frequencies using the extension assay described above in a group of samples isolated from 92 individuals originating from the state of Utah in the United States, Venezuela and France (Coriell cell repositories).

[0247] Genotyping results are shown for female pools in Table 6A and 6B. Table 6A shows the orginal genotyping results and Table 6B shows the genotyped results re-analyzed to remove duplicate individuals from the cases and controls (*i.e.*, individuals who were erroneously included more than once as either cases or controls). Therefore, Table 6B represents a more accurate measure of the allele frequencies for this particular SNP. In the subsequent tables, "AF" refers to allelic frequency; and "F case" and "F control" refer to female case and female control groups, respectively.

TABLE 6A

Reference SNP ID	AF F case	AF F control	p-value	Breast Cancer Assoc. Allele
rs1671152	T = 0.139 G = 0.861	T = 0.192 G = 0.808	**0.0196**	G
rs1050348	C = 0.625 T = 0.375	C = 0.540 T = 0.460	**0.0050**	C
rs454422	C = 0.831 A = 0.169	C = 0.761 A = 0.239	**0.0049**	C
rs763471	T = 0.523 G = 0.477	T = 0.593 G = 0.407	**0.0079**	G
rs2046778	A = 0.836 G = 0.164	A = 0.736 G = 0.264	**0.0019**	A

TABLE 6B

Reference SNP ID	AF F case	AF F control	p-value	Odds Ratio	Breast Cancer Assoc. Allele
rs1671152	T = 0.143 G = 0.857	T = 0.190 G = 0.810	**0.00109**	0.65	G (T323K)
rs1050348	C = 0.629 T = 0.371	C = 0.543 T = 0.457	**0.0124**	0.72	C (H491Y)
rs454422	C = 0.834 A = 0.166	C = 0.762 A = 0.238	**0.00452**	1.57	C
rs763471	T = 0.520 G = 0.480	T = 0.596 G = 0.404	**0.0166**	0.74	G
rs2046778	A = 0.811 G = 0.189	A = 0.724 G = 0.276	**0.00114**	0.61	A

[0248] The single marker alleles set forth in Table 3 were considered validated, since the genotyping data for the females, males or both pools were significantly associated with breast cancer, and because the genotyping results agreed with the original allelotyping results. Particularly significant associations with breast cancer are indicated by a calculated p-value of less than 0.05 for genotype results, which are set forth in bold text. Tables 6A and 6B show the disease associated allele in column 6. In the case of rs1671152, this SNP is an exonic SNP that codes for a T323K amino acid change in the GP6 gene. The guanine allele codes for threonine (T); therefore, a threonine is associated with an increased risk of breast cancer. In the case of rs454422, this SNP is an exonic SNP that codes for a H491Y amino acid change in the LAMA4 gene. The cytosine allele codes for histidine (H); therefore, a histidine is associated with an increased risk of breast cancer.

[0249] Odds ratio results are shown in Tables 6B. An odds ratio is an unbiased estimate of relative risk which can be obtained from most case-control studies. Relative risk (RR) is an estimate of the likelihood of disease in the exposed group (susceptibility allele or genotype carriers) compared to the unexposed group (not carriers). It can be calculated by the following equation:

RR = *I*A/*I*a

*I*A is the incidence of disease in the A carriers and *I*a is the incidence of disease in the non-carriers.

RR > 1 indicates the A allele increases disease susceptibility.

RR < 1 indicates the a allele increases disease susceptibility.

For example, RR = 1.5 indicates that carriers of the A allele have 1.5 times the risk of disease than non-carriers, *i.e.*, 50% more likely to get the disease.

[0250] Case-control studies do not allow the direct estimation of *I*A and *I*a, therefore relative risk cannot be directly estimated. However, the odds ratio (OR) can be calculated using the following equation:

OR = (nDAnda)/(ndAnDa) = *p*DA(1 – pdA)/*p*dA(1 – pDA), or

OR = ((case f) / (1- case f)) / ((control f) / (1-control f)), where f = susceptibility allele frequency.

[0251] An odds ratio can be interpreted in the same way a relative risk is interpreted and can be directly estimated using the data from case-control studies, *i.e.*, case and control allele frequencies. The higher the odds ratio value, the larger the effect that particular allele has on the development of breast cancer. Possessing an allele associated with a relatively high odds ratio translates to having a higher risk of developing or having breast cancer.

Example 3

GP6 Region Proximal SNPs

[0252] It has been discovered that a polymorphic variation (rs1671152) in a region that encodes the glycoprotein VI (platelet) (GP6) gene is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs1671152) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. Approximately 124 allelic variants located within or near the GP6 gene were identified and 114 SNPs were allelotyped. The polymorphic variants are set forth in Table 7. The chromosome position provided in column four of Table 7 is based on Genome "Build 33" of NCBI's GenBank.

TABLE 7

dbSNP rs#	Chromosome	Position in Figure 1	Chromosome Position	Allele Variants
269911	19	185	60156885	A/T
703464	19	237	60156937	G/A
703465	19	641	60157341	C/G
269912	19	719	60157419	A/G
269913	19	990	60157690	T/C
269915	19	2908	60159608	C/T
269916	19	3140	60159840	G/A
172006	19	3880	60160580	A/G
703467	19	4494	60161194	C/T
703468	19	5107	60161807	G/A
2217659	19	5220	60161920	G/A
775894	19	6031	60162731	A/G
775900	19	8670	60165370	A/G
1654491	19	13794	60170494	T/C
775903	19	16356	60173056	A/G
1036231	19	17164	60173864	T/C
1036232	19	17264	60173964	G/A
1671223	19	20537	60177237	T/G
1671224	19	20637	60177337	A/G
1654495	19	20900	60177600	C/A
1654496	19	21155	60177855	C/A
1654497	19	21795	60178495	T/C
1671225	19	21931	60178631	T/G
1671226	19	22167	60178867	C/G
1671227	19	22656	60179356	T/C
1654498	19	23108	60179808	T/C
1671228	19	23404	60180104	T/C
1654499	19	24287	60180987	T/C
1869616	19	24480	60181180	A/C
1654503	19	24592	60181292	C/T

dbSNP rs#	Chromosome	Position in Figure 1	Chromosome Position	Allele Variants
1654504	19	24878	60181578	T/C
1671133	19	26370	60183070	A/C
1654505	19	27056	60183756	G/A
3786863	19	27874	60184574	A/G
1560714	19	31248	60187948	G/A
1654406	19	31458	60188158	G/T
1043673	19	31553	60188253	C/A
1043678	19	31637	60188337	G/T
1043680	19	31668	60188368	C/G
1043684	19	31752	60188452	A/G
1671140	19	37643	60194343	G/A
1654409	19	43941	60200641	A/C
1654410	19	44134	60200834	T/C
1654411	19	44329	60201029	A/C
1671148	19	44343	60201043	A/C
1671149	19	44362	60201062	G/A
1671150	19	44818	60201518	G/A
1654412	19	44917	60201617	C/T
1671151	19	45215	60201915	G/A
1671152	19	45666	60202366	T/G
1654413	19	45680	60202380	T/A
2304167	19	46402	60203102	C/T
1671153	19	46510	60203210	G/T
2019599	19	46554	60203254	C/T
1654485	19	46823	60203523	A/C
1671188	19	47714	60204414	T/G
1671191	19	48963	60205663	T/C
1654415	19	49157	60205857	C/T
1671192	19	49254	60205954	G/A
2304168	19	49257	60205957	A/G
1654416	19	49356	60206056	T/C
1654419	19	55202	60211902	A/G
1654420	19	55527	60212227	T/A
1613662	19	55916	60212616	G/A
2886415	19	56402	60213102	T/G
2365593	19	56413	60213113	C/T
2886414	19	56685	60213385	G/A
1654421	19	56783	60213483	A/G
1654424	19	58044	60214744	A/G
1654425	19	58301	60215001	T/C
892089	19	58382	60215082	A/G
892090	19	58393	60215093	G/T
1671196	19	58869	60215569	C/T
1671198	19	59155	60215855	T/C
1671199	19	59189	60215889	G/A
1625609	19	62546	60219246	C/T
1625689	19	62568	60219268	G/A

dbSNP rs#	Chromosome	Position in Figure 1	Chromosome Position	Allele Variants
1654438	19	70983	60227683	A/G
2569513	19	71465	60228165	G/A
2569514	19	71538	60228238	G/A
1671214	19	72144	60228844	G/A
1671215	19	72340	60229040	C/A
1054796	19	72527	60229227	C/G
1654439	19	72968	60229668	T/G
1671216	19	73397	60230097	A/G
1671217	19	73553	60230253	G/A
1671218	19	73720	60230420	C/T
1654441	19	74190	60230890	C/T
1654442	19	74687	60231387	T/G
1671219	19	74699	60231399	G/A
10666	19	75580	60232280	C/T
1626971	19	76345	60233045	T/C
1671221	19	76506	60233206	G/A
754235	19	77554	60234254	G/A
775821	19	77889	60234589	C/T
3745912	19	77919	60234619	A/G
775822	19	78866	60235566	A/G
1059211	19	79061	60235761	C/T
2124090	19	83777	60240477	A/C
1671171	19	84360	60241060	T/G
1671170	19	84631	60241331	T/A
1654444	19	85775	60242475	G/T
2365721	19	87153	60243853	G/A
1654446	19	89650	60246350	A/G
1654447	19	89895	60246595	A/G
1671176	19	90103	60246803	C/A
1654448	19	90234	60246934	G/A
1671178	19	90309	60247009	G/A
1654449	19	90376	60247076	G/A
1671182	19	90925	60247625	C/T
1654451	19	91561	60248261	A/T
1654452	19	91605	60248305	G/A
1671169	19	92954	60249654	T/C
1654459	19	94228	60250928	A/G
269909	19	Not Mapped		G/C
269910	19	Not Mapped		T/G
776251	19	Not Mapped		G/A
892088	19	Not Mapped		A/G
892091	19	Not Mapped		C/G
1043680	19	Not Mapped		C/G
1064675	19	Not Mapped		A/G
1671187	19	Not Mapped		T/A
2116883	19	Not Mapped		T/C
2163833	19	Not Mapped		G/A

Assay for Verifying and Allelotyping SNPs

[0253] The methods used to verify and allelotype the proximal SNPs of Table 7 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 8 and Table 9, respectively.

TABLE 8

dbSNP rs#	Forward PCR primer	Reverse PCR primer
10666	ACGTTGGATGAGATGGCCCCTCTCCCCT	ACGTTGGATGAGTGGACTGGCCTGCAGGT
172006	ACGTTGGATGGGTTGAGGAGTATTCCATTG	ACGTTGGATGTGGGTGACAGCAAGACTCCA
269909	ACGTTGGATGAGTTTCTTGTCTCCTGGGTG	ACGTTGGATGCACAACAATGAAAGGACAAGC
269910	ACGTTGGATGGGCTGCTCAGGTTCTAAAAG	ACGTTGGATGCCCCAGTTCCTTATCTGATC
269911	ACGTTGGATGGGTGACAAAGTGAGACTCCG	ACGTTGGATGTGACTAGCTGGGATTATGGG
269912	ACGTTGGATGAGGAGAGCCTGCAGGTTGAA	ACGTTGGATGTCCCTTGATTGTCATCACAG
269913	ACGTTGGATGAGATGCACCGAATGGATCTG	ACGTTGGATGTCCTAGGACACAGTGTGGAC
269915	ACGTTGGATGAAGCTGAGATTGTGCTGCTG	ACGTTGGATGACTACTCTACTTTCTACCCC
269916	ACGTTGGATGAACTCCTGACCACGTGATCC	ACGTTGGATGAAAGTGTAGCTGGGCATGGC
703464	ACGTTGGATGAAAAGTAGCTGAGCGTGGTG	ACGTTGGATGCCCAGGTTCAAACAGTTCTC
703465	ACGTTGGATGGCTGTGATGACAATCAAGGG	ACGTTGGATGAGCAGGAACTGGCATTTGAG
703467	ACGTTGGATGAAACTTACGAAGGTCTGGGC	ACGTTGGATGGGGTTTCAGGAACATTCACC
703468	ACGTTGGATGGAAAGGAACAAGTGATCCAG	ACGTTGGATGCTTCTGAAAAAGGAGAAGGG
754235	ACGTTGGATGGTGGAACAAACAGTTGGAGC	ACGTTGGATGGAGTGGAGATCATGCCATTG
775821	ACGTTGGATGTTTTCTCATCCGCCAGCAGG	ACGTTGGATGACAGAGCACAGGTCCCTTTC
775822	ACGTTGGATGAAAAGTAGAGCATGTGGACC	ACGTTGGATGGGGATGAACAGACAATCCTC
775894	ACGTTGGATGAGGCAGGAGGACTGATTGAG	ACGTTGGATGTTGTAGAGACAGGGTCTTGC
775900	ACGTTGGATGTCTGCAACGTTCGTTCTTCC	ACGTTGGATGTTCTTAATGTGCCCACTGTC
775903	ACGTTGGATGTGTGCCCGGCCCTTTTTTTC	ACGTTGGATGGGACTCCAGTCTGAGTGACA
776251	ACGTTGGATGAATCCTAGCTACTCAGGAGG	ACGTTGGATGATGGTATGATCTAGGCTCCC
892088	ACGTTGGATGATTGGTTACACTGGGACTGC	ACGTTGGATGAAGCGGTGATTCTCAGCCTC
892089	ACGTTGGATGCCCTACAAGAATCCCGAGAG	ACGTTGGATGAAGCTGTAGCATCGGTAGGTT
892090	ACGTTGGATGAAGCTGTAGCATCGGTAGGTT	ACGTTGGATGCCCTACAAGAATCCCGAGAG
892091	ACGTTGGATGAAGATGGAGGACCACAGGTG	ACGTTGGATGACCGGAATTACTGAGAGGTC
1036231	ACGTTGGATGAACAGTACTAAGGGCAGATG	ACGTTGGATGGTCCAGGGTGTTTACTGTTC
1036232	ACGTTGGATGCCAAAGAAACTCCCAGAATC	ACGTTGGATGGATCGTGCCATTGCACTTTG
1043673	ACGTTGGATGTGAAGTCATGAGAAGAAGGC	ACGTTGGATGGCTGCTGGAAGAAATAGAAG
1043678	ACGTTGGATGTTCATGATCTGAATCCCCCC	ACGTTGGATGGAACTCATTATCCCTGAGGG
1043680	ACGTTGGATGTCTAGCCCAGCAATGAACTC	ACGTTGGATGTTCCAGGTGTTGGTGAACTG
1043684	ACGTTGGATGCTCAATATACCCGTGATACAG	ACGTTGGATGTTTAGCCATGATTCTGCCTC
1054796	ACGTTGGATGAGTGTCACCCTGATTTCCAG	ACGTTGGATGCACCTTGGGAAATACGTTGC
1059211	ACGTTGGATGCTCAGCTCCTTGGTGAAGAG	ACGTTGGATGGGCAGACGAGGAAGTATAAC
1064675	ACGTTGGATGGAGTTCCCTCAGTTTTTATTG	ACGTTGGATGCCTACTACATTCCTTTTTGC
1560714	ACGTTGGATGATCTGCTGACCTCGTGATCC	ACGTTGGATGAAAAGACAGTCTCAGGTGGG
1613662	ACGTTGGATGATGGACCCTGCAGAACCTAC	ACGTTGGATGTCTGATTTCCCAGGAACCTC
1625609	ACGTTGGATGTCAAGCGATTCTCCTGCCTC	ACGTTGGATGAAAAAATGAGCTGGGCGTGG
1625689	ACGTTGGATGGTAATCCCAGCTACTTGGAG	ACGTTGGATGATCTTGGCTCACTGCAGCCT
1626971	ACGTTGGATGTATTAAATGCACCTGGCACC	ACGTTGGATGCAAAGTGCTGGGATTACAGG

dbSNP rs#	Forward PCR primer	Reverse PCR primer
1654406	ACGTTGGATGCTTCTATTTCTTCCAGCAGC	ACGTTGGATGTTTCTTCCCCCATTGTACCC
1654409	ACGTTGGATGGTGAAACCTTGTCTCATATAC	ACGTTGGATGTGAGAGTAGGCATGTGGTAC
1654410	ACGTTGGATGACTGTGCCTAGGCTATACTG	ACGTTGGATGGGAAAATCATACTGAGATGC
1654411	ACGTTGGATGTGGAACTTCTTGGTGCCATC	ACGTTGGATGCCTGTAATCCCGGCACTTTG
1654412	ACGTTGGATGATTAGCCAGGTGTGGTGGTG	ACGTTGGATGTCAAGCCATTCTCCCACCTC
1654413	ACGTTGGATGAGGGCTGTGCAGAGGCCGCTT	ACGTTGGATGTCCATCCTGACCCCCGTTTG
1654415	ACGTTGGATGCCAAGAAAGTCCTTGGTGTG	ACGTTGGATGCTTTGAAATGGCCCCATCAC
1654416	ACGTTGGATGTCTGCTGAGCATGAAATGCC	ACGTTGGATGCTGAACTGACCGTCTCATTC
1654419	ACGTTGGATGTATCATACGCTAGGCTGGAG	ACGTTGGATGATGTTTCTCCTGCCTTGGTG
1654420	ACGTTGGATGCCAACCAACCAACAAACCTG	ACGTTGGATGTGGAAGTTTGAGAACCGCTG
1654421	ACGTTGGATGAGGACACAGGAATCCAGAAG	ACGTTGGATGGCACATTCTGGGCTATTAAC
1654424	ACGTTGGATGTAGGTGGGAAGGAAGTGGGA	ACGTTGGATGCCACTTCTTTCCCACCTATG
1654425	ACGTTGGATGTACCTGTGACCACAAGCTCC	ACGTTGGATGTGCTACAGCTTCTCCAGCAG
1654438	ACGTTGGATGAATCAACTAGGCATGGTGGC	ACGTTGGATGCCAGGTTCAAGCGATTCTCC
1654439	ACGTTGGATGCCCCATATACATGTGCGATG	ACGTTGGATGAATGGGGTGTTTCTGGAGCA
1654441	ACGTTGGATGAGTAGCTGGGATTACAGGCG	ACGTTGGATGGGAGTTCAAGATAAGCCTGG
1654442	ACGTTGGATGAGGAGAATGGTGTGAAGCTG	ACGTTGGATGAATCTTGCTCTGTCACCCAG
1654444	ACGTTGGATGGGATGGTCCCAGTTTTACAT	ACGTTGGATGCCAGGAGAATCACTTTTATGG
1654446	ACGTTGGATGAAAAGGAAGGGCATTCTGGC	ACGTTGGATGTTTGGCCTCCCAAAGTACTG
1654447	ACGTTGGATGATCCCTGGGAAGACGGTCAT	ACGTTGGATGTTACCTCTCCTGGCCAGTTC
1654448	ACGTTGGATGTGCTCACTGCATGAGATTCC	ACGTTGGATGAACTTTGGCCTCCCAAAGTG
1654449	ACGTTGGATGAGTCCAGCCTGGCAAACATG	ACGTTGGATGCAGTCTAATCTCTCTTTTCCC
1654451	ACGTTGGATGTTTAAAATGCCCGCTGCACG	ACGTTGGATGAGGAGGATGCACTTATGTGG
1654452	ACGTTGGATGCTGTACGCATTACCACAGAC	ACGTTGGATGGTTTTGGACTCTTGACCTGC
1654459	ACGTTGGATGCAGGAGCTTGGGTACCCAC	ACGTTGGATGCCCTCATCTGGAAATGTGTG
1654485	ACGTTGGATGTTGTACCACTGCACTCTAGC	ACGTTGGATGCCTGACTCTACAGTTCTTGC
1654491	ACGTTGGATGCAGACGTCCGTGCTTCACC	ACGTTGGATGTCCAGGAACAGACGGAGGTC
1654495	ACGTTGGATGATGACCATTGCTCGTCTGTG	ACGTTGGATGGCTTTCTGCAGAGGTTGTCG
1654496	ACGTTGGATGAATCACAAATGGCAACACGG	ACGTTGGATGTTTGGATGCTGGCACTTGTG
1654497	ACGTTGGATGACCCCATGCTGTGTTTTCTC	ACGTTGGATGCAGAAGACTACCTGATTTGC
1654498	ACGTTGGATGCTTCCCACACCCACTATATC	ACGTTGGATGGTTAGTGAGTCGGTGACATC
1654499	ACGTTGGATGCACTACCTCTCTAGCAACTG	ACGTTGGATGACCTCAGATGATCTGCCCAC
1654503	ACGTTGGATGTCCTTGGCTTGTGGCCCTTC	ACGTTGGATGAGCCAGGGCAACGTTTGAAG
1654504	ACGTTGGATGCCACCCCATGATTCCATTTC	ACGTTGGATGTGCTGTGATGCACCTTTGAC
1654505	ACGTTGGATGCCCTGTCTCTCTAAAACCAC	ACGTTGGATGATTCAAGCAGTTCTCGTGCC
1671133	ACGTTGGATGGTGGTCTCAACTTGGCTATC	ACGTTGGATGCCAGATAGGATTCCAGGTTC
1671140	ACGTTGGATGAGTCTGACAAGAGAGTCAGC	ACGTTGGATGTCCTTTACCTACCCACATCC
1671148	ACGTTGGATGGCCATCCTTCTGTCTTTTCC	ACGTTGGATGAGTGGCTCATGCCTGTAATC
1671149	ACGTTGGATGCTTTTCCCAAGTGACTCACC	ACGTTGGATGAAAAGAATGGCTGGCCACAG
1671150	ACGTTGGATGGTGCTATGATCAAATCAGGG	ACGTTGGATGACACCACTGCACTCTAGCTC
1671151	ACGTTGGATGGGAAAACCAGACAAGAGCAC	ACGTTGGATGTGACTCTGTTCCATCCTCTG
1671152	ACGTTGGATGAGGGCTGTGCAGAGGCCGCTT	ACGTTGGATGTGAACATCCTGTCGGCCTCC
1671153	ACGTTGGATGCCTACTCCGAACACACACAC	ACGTTGGATGATTATAGGCATGAGGCACCG
1671169	ACGTTGGATGTCCTGTTGCTGGACACTATC	ACGTTGGATGTCACACCTTCCGAGGATTAG
1671170	ACGTTGGATGAGGTGACAGTGCTGTACCTG	ACGTTGGATGACAAAGAACAGTGAGAGGGC
1671171	ACGTTGGATGAAGCAAGATACCGTCTCAGA	ACGTTGGATGCCGGGAAATGGAATAATTCC
1671176	ACGTTGGATGTGGAGCCACTTATGGAGAAC	ACGTTGGATGACCCCAACTGAAACACAGAC

dbSNP rs#	Forward PCR primer	Reverse PCR primer
1671178	ACGTTGGATGTAATCCCAGCACTTTGGGAG	ACGTTGGATGCATGTTTGCCAGGCTGGACT
1671182	ACGTTGGATGATAGGGCGGCTTTTCTCCTG	ACGTTGGATGCCTGGGAACTGAATGTCTCG
1671187	ACGTTGGATGAGTGCTCAGCAACGATTACG	ACGTTGGATGGAGGGCTGCAGGTTGAGAAA
1671188	ACGTTGGATGGGAACCGCAGATGGACAATG	ACGTTGGATGAGATCACAGAGTGAGGAGAG
1671191	ACGTTGGATGTCGGACGCACACAGACTGTAG	ACGTTGGATGGGAAAGCGTATCTGCAGAGG
1671192	ACGTTGGATGTGGTAAGAGACGGACAGTTC	ACGTTGGATGTCAGCAGAAAGGAGTGTGAG
1671196	ACGTTGGATGTTGCTAGGCAACAGGCACTC	ACGTTGGATGTCTGTATCTGAGCCTCACTG
1671198	ACGTTGGATGATGAAACTAAGGCACATGGC	ACGTTGGATGCTTATAATCTACCCTCTTAGC
1671199	ACGTTGGATGGCTGAAATTTGCTAAGAGGG	ACGTTGGATGGACAGTTACTACTAGCAAGC
1671214	ACGTTGGATGAGGCGGAGAATGATCCGGTG	ACGTTGGATGACGCCATCATTCGTGCATCC
1671215	ACGTTGGATGTTCTCCAAAGCACCCAAGTG	ACGTTGGATGATGCTGGGCTTGCTTTTTCC
1671216	ACGTTGGATGTGCTTGGGAGCAAGTTACAG	ACGTTGGATGTTCCCCCTCCTGGTATTTAC
1671217	ACGTTGGATGTTGTCTCCATTCCTCCCTGG	ACGTTGGATGTCTTGTCTTGCCCTCTCGCT
1671218	ACGTTGGATGTGAGTCTGGTAGGCAACTTC	ACGTTGGATGTAGAAGCCAGTCGCTACATC
1671219	ACGTTGGATGTGATCTCGGCTCACTGCAAG	ACGTTGGATGAAATTAGCTGGGCATGGTGG
1671221	ACGTTGGATGTGGTGAAACCCCATTTCTAC	ACGTTGGATGGGTTCAAGGGATTCTCCTGC
1671223	ACGTTGGATGTCAAGTGATTCTCCTGCCTC	ACGTTGGATGCCCACCTCTACTGAAAATAC
1671224	ACGTTGGATGTGAGTCTCACTCTTGTTGCC	ACGTTGGATGCAGGAGAATCACTTGAACCC
1671225	ACGTTGGATGTATAGGCGTGAGCCACTATG	ACGTTGGATGCTATTGGAAGCTACATGCTC
1671226	ACGTTGGATGTATTGGCCAGACTGGACTTC	ACGTTGGATGAGTTACTCAGGAGGCTAAGG
1671227	ACGTTGGATGGGTTTCTGTTCAGAGATTCG	ACGTTGGATGTGCAGTGAGCCTAGATCATG
1671228	ACGTTGGATGTCAGCCTCCCAGGGATTAAG	ACGTTGGATGACATGGTGAAAACTCGTCTC
1869616	ACGTTGGATGTAATCCCAGCTACTCGGAAG	ACGTTGGATGACGGTGGCTCACTTCAACCT
2019599	ACGTTGGATGGTGCTGGGATTATAGGCATG	ACGTTGGATGTACTCCGAACACACACACAC
2116883	ACGTTGGATGATTACAGGCATGAGCCACTG	ACGTTGGATGCACGCGCAGTTCAATTTCTC
2124090	ACGTTGGATGTCTGACAAAGCTGGAAGCTG	ACGTTGGATGCTGATAAACAAGGCTGTGGG
2163833	ACGTTGGATGGATATTGGTGAGTATGCAGAG	ACGTTGGATGAACTGTTTTCCACAGCAGGG
2217659	ACGTTGGATGTTCCCCCCTTCTCCTTTTTC	ACGTTGGATGATGAGGTAACTTACCTAATG
2304167	ACGTTGGATGGTTTGGTTCCCAGAGACTTC	ACGTTGGATGAGGATGACTTACTCACCAGC
2304168	ACGTTGGATGTCAGCAGAAAGGAGTGTGAG	ACGTTGGATGTGGTAAGAGACGGACAGTTC
2365593	ACGTTGGATGTGACGCAGTAAGACTCCATC	ACGTTGGATGCAAAGTGCTGGGATTACAGG
2365721	ACGTTGGATGTTGTACAGCCTGCAAGCAAC	ACGTTGGATGAGATCGCGCCATTGCACTCA
2569513	ACGTTGGATGGTTGGCGTTTTTGTTTGCAC	ACGTTGGATGTCTCATAGTATTCTGCAGGG
2569514	ACGTTGGATGTCCCTGCAGAATACTATGAG	ACGTTGGATGAGAGTGTTGGGATTACAGGC
2886414	ACGTTGGATGGGTGTGCTTTACAAATGCTG	ACGTTGGATGAACTGAGATCACTCCACTGC
2886415	ACGTTGGATGTGACGCAGTAAGACTCCATC	ACGTTGGATGCAAAGTGCTGGGATTACAGG
3745912	ACGTTGGATGACGTCTTCTGAGGCACAGAG	ACGTTGGATGGCTGTTAGAGGCTGGCAGG
3786863	ACGTTGGATGTGACCAACAGAAGTCTCAGG	ACGTTGGATGTTGACCTCAGGTGATCCATC

TABLE 9

dbSNP rs#	Extend Primer	Term Mix
10666	TGCAGGTGAGCACTGCCC	ACG

dbSNP rs#	Extend Primer	Term Mix
172006	GCAAGACTCCATCTCAA	ACT
269909	AAGCATAGATCAGATAAGGAA	ACT
269910	ATCTATGCTTGTCCTTTCAT	ACT
269911	GCTCAGCTACTTTTTGTAT	CGT
269912	CAAGATGGTGTCTTCGGC	ACT
269913	ACAGTGTGGACCGATTTCC	ACT
269915	AGACAAGTCTCACTCTG	ACG
269916	GGCGGCTCACACCTGTAAT	ACG
703464	CTCCTGCCTCAGCCTCC	ACG
703465	TGGCATTTGAGACAGGA	ACT
703467	CATTCACCATGTCTGTGTGAG	ACG
703468	CTTCATAAAAGAAAAGATGACA	ACG
754235	CATGCCATTGTACTCCAGCC	ACG
775821	CCTGCCAGCCTCTAACAGC	ACG
775822	TAGTGATGTCTGCTTCAG	ACT
775894	TTGCCCAGGCTGGCCTC	ACT
775900	GAATGCCAACCTCCCTTCC	ACT
775903	TCTGAGTGACAGAGCGA	ACT
776251	GCTCCCCGCAACCTCCGC	ACG
892088	GCCTCGGCCGCAATCACA	ACT
892089	CGGTCACCGTGATGATGGG	ACT
892090	AGAATCCCGAGAGATGGTAC	CGT
892091	GTCCTTCACCTGAGCTTCC	ACT
1036231	GTGTTTACTGTTCAAGGCAAGT	ACT
1036232	GGCAACAGAGCAAGACT	ACG
1043673	ACTGAGAAACATCATCCCTGGG	CGT
1043678	AGTCACAGGCAGTTCACC	CGT
1043680	CTGTGACTCCTCTCCTCCCC	ACT
1043684	CTGTTTTATACCTGCACAC	ACT
1054796	ACGCCAGGCAGGCTCTCA	ACT
1059211	CGCCTACTGCCAGAGCAAGCT	ACG
1064675	ATTCCTTTTTGCTGAAATAATGAA	ACT
1560714	TGGGGCGTGATGGCTCA	ACG
1613662	CAACAGAACCACCTTCC	ACG
1625609	GTGCACACCTGTAATCC	ACG
1625689	CAGGGCTCAAGCGATTCTCC	ACG
1626971	TCGCCTGGCCAAAAAAA	ACT
1654406	CATTGTACCCCAGGTTGAAAAT	CGT
1654409	GTGGTACCACACCCAGCTAATT	ACT
1654410	ATCATACTGAGATGCTATCAGAA	ACT
1654411	GCACTTTGGGAGGTTGAGG	ACT
1654412	CATTCTCCCACCTCAGCCCCC	ACG
1654413	CCCGTTTGATTTCCGGGTC	CGT
1654415	GGCCCCATCACCCAAAA	ACG

dbSNP rs#	Extend Primer	Term Mix
1654416	GACCGTCTCATTCACAAAC	ACT
1654419	TTGGTGCTTCACTCTGAGAC	ACT
1654420	GAGAACCGCTGATCAATGCA	CGT
1654421	GCATGCAGCTCCCGTCC	ACT
1654424	CCACCTATGGCCGCGCCCCT	ACT
1654425	CAGGGACCCATACCTGTGGTC	ACT
1654438	TCAGCCTCCTGAGTAGCTGG	ACT
1654439	GTTTCTGGAGCACTCCGGT	ACT
1654441	GATAAGCCTGGCCAACA	ACG
1654442	ATGATCTCGGCTCACTGCAA	ACT
1654444	TATGGATCTTTCTAGTCTTGTTT	CGT
1654446	ACTGATTACAGGCGTGC	ACT
1654447	CCCGATGCCTGTGTTGGC	ACT
1654448	AGTGCTGGGATTACAGG	ACG
1654449	AATCTCTCTTTTCCCTACACA	ACG
1654451	TAATGCGTACAGCAGCC	CGT
1654452	ACTGGAGGAGGATGCACTTA	ACG
1654459	ATGCACAGAAACAAGGATCTA	ACT
1654485	CTTGCTTTTTTTTTTTTGGACAG	ACT
1654491	GCACCCCGAGCCTTTCCAG	ACT
1654495	TTGTCGTAAGTCTCTCCTCTCTT	CGT
1654496	CGGGAAGGTTGAAGTTGGAC	CGT
1654497	CCATTTACAACCAATTGC	ACT
1654498	CTTGTGGGACTTCTTTTTTA	ACT
1654499	ACCCTGGCCTCCCTAAC	ACT
1654503	GGGCAACGTTTGAAGATGCTCTGC	ACG
1654504	CACCTTTGACTCTTGAGCC	ACT
1654505	TAGCTATGTGCCACCATGCC	ACG
1671133	GATTGTAGCTAACTCACAAGG	ACT
1671140	TACCTACCCACATCCTATAAAA	ACG
1671148	CCTGTAATCCCGGCACT	ACT
1671149	CTGGCCACAGTGGCTCA	ACG
1671150	CGGGTGACGAAGCCTGAC	ACG
1671151	TCCTCTGTGCAAAATCCTCC	ACG
1671152	CTCCATCCTGACCCCCGT	ACT
1671153	CTGTGGAATTGTGCCTC	CGT
1671169	CATGTCCCACAGAGGCTAAC	ACT
1671170	GAGAGGGCAATGCCTCAGAG	CGT
1671171	TTCTGGGATTCTCTAGAGGG	ACT
1671176	AGACATCATCACATCACACCA	CGT
1671178	CCAGGCTGGACTCGAACT	ACG
1671182	ACTGAATGTCTCGGTATAAAACC	ACG
1671187	CAGGTTGAGAAAGCTCTA	CGT
1671188	CAGAGTGAGGAGAGTGAGAC	ACT

dbSNP rs#	Extend Primer	Term Mix
1671191	GAGCGGTTAGAAGATGTGCT	ACT
1671192	AAGCCTGTAGGCTTTTAA	ACG
1671196	GGGATGACTGAATGAGACAGTA	ACG
1671198	CCCTCTTAGCAAATTTCAGCT	ACT
1671199	TAACTTTTTTGTGTGTGAGAA	ACG
1671214	CGTGCATCCTTCCCACCTA	ACG
1671215	GTACTCAAGATGATGTAA	CGT
1671216	TTACACCCTGGAGTGGTCC	ACT
1671217	TGCCCTCTCGCTGGCTGG	ACG
1671218	CAAAGGGAGGTGGTCGCAC	ACG
1671219	CAGGAGAATGGTGTGAACC	ACG
1671221	AGCTGGGATTACAGGCA	ACG
1671223	TACAAAATTAGCTGGGCATG	ACT
1671224	CTGTGAGCCGAGATTGC	ACT
1671225	CTCAATGTGATCCTCCT	ACT
1671226	GCAGGAGAATCACTTGAACTT	ACT
1671227	AGATCATGCCATTGCCAGC	ACT
1671228	ACAGAAGTTAGCTGGGC	ACT
1869616	CTTCAACCTCCGCCTCCTGG	ACT
2019599	GAAAAGCATGGGCCGGGCA	ACG
2116883	CATACTCACCAATATCTGCT	ACT
2124090	GCTTTGTGTTCTTTCTAGTC	ACT
2163833	GCCAGCAATGCACGCGCAGT	ACG
2217659	GTAACTTACCTAATGATAGAGG	ACG
2304167	TGACTCCTTTGGACTGG	ACG
2304168	CAGTTCGGTGAAGTGGTT	ACT
2365593	GGTGTGAGCCACCACGCC	ACG
2365721	AGACTCCCTCTCAAAATAA	ACG
2569513	AGGGATAAGCATGAAACCACT	ACG
2569514	GCGTGAGCCACCACGCC	ACG
2886414	GGGTGACAAAGTGAGACTC	ACG
2886415	CACGCCTGGCTAAGCCT	ACT
3745912	GGCTGGCAGGCCAGGTCAAC	ACT
3786863	GTGCTGGGATTACAGGC	ACT

Genetic Analysis of Allelotyping Results

[0254] Allelotyping results are shown for cases and controls in Table 10. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where "AF" is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP

rs269911has the following case and control allele frequencies: case A1 (A) = 0.907; case A2 (T) = 0.093; control A1 (A) = 0.864; and control A2 (T) = 0.136, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped.

TABLE 10

dbSNP rs#	Position in Figure 1	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
269911	185	60156885	A/T	0.136	0.093	0.0242
703464	237	60156937	G/A	0.219	0.197	0.3696
703465	641	60157341	C/G	0.608	0.634	0.3828
269912	719	60157419	A/G	0.621	0.651	0.3011
269913	990	60157690	T/C	0.354	0.320	0.2275
269915	2908	60159608	C/T	0.776	0.797	0.3802
269916	3140	60159840	G/A	0.137	0.123	0.5037
172006	3880	60160580	A/G	0.172	0.142	0.1703
703467	4494	60161194	C/T	0.321	0.273	0.0759
703468	5107	60161807	G/A	0.187	0.159	0.2093
2217659	5220	60161920	G/A	0.167	0.127	0.0593
775894	6031	60162731	A/G	0.152	0.131	0.3098
775900	8670	60165370	A/G	0.632	0.667	0.2304
1654491	13794	60170494	T/C	0.005	0.005	0.9971
775903	16356	60173056	A/G	0.818	0.835	0.4460
1036231	17164	60173864	T/C	0.177	0.178	0.9755
1036232	17264	60173964	G/A	0.690	0.713	0.3949
1671223	20537	60177237	T/G	0.951	0.978	0.0143
1671224	20637	60177337	A/G	0.235	0.244	0.7269
1654495	20900	60177600	C/A	0.534	0.536	0.9379
1654496	21155	60177855	C/A	0.315	0.289	0.3567
1654497	21795	60178495	T/C	0.307	0.315	0.7771
1671225	21931	60178631	T/G	0.352	0.344	0.7785
1671226	22167	60178867	C/G	0.520	0.482	0.2147
1671227	22656	60179356	T/C	0.468	0.463	0.8612
1654498	23108	60179808	T/C	0.385	0.387	0.9372
1671228	23404	60180104	T/C	0.412	0.403	0.7790
1654499	24287	60180987	T/C	0.349	0.355	0.8197
1869616	24480	60181180	A/C	0.909	0.925	0.3271
1654503	24592	60181292	C/T	0.384	0.409	0.3829
1654504	24878	60181578	T/C	0.689	0.651	0.1817
1671133	26370	60183070	A/C	0.345	0.348	0.9023
1654505	27056	60183756	G/A	0.257	0.263	0.8407
3786863	27874	60184574	A/G	0.989	0.989	0.9843
1560714	31248	60187948	G/A	0.970	0.981	0.2419
1654406	31458	60188158	G/T	0.437	0.437	0.9994
1043673	31553	60188253	C/A	0.354	0.374	0.4936
1043678	31637	60188337	G/T	0.530	0.525	0.8676
1043680	31668	60188368	C/G	0.544	0.561	0.5849
1043684	31752	60188452	A/G	0.317	0.312	0.8472
1671140	37643	60194343	G/A	0.122	0.116	0.7440
1654409	43941	60200641	A/C	0.557	0.528	0.3425
1654410	44134	60200834	T/C	0.431	0.462	0.3094
1654411	44329	60201029	A/C	0.639	0.578	0.0398
1671148	44343	60201043	A/C	0.859	0.824	0.1102
1671149	44362	60201062	G/A	0.798	0.745	0.0377
1671150	44818	60201518	G/A	0.116	0.148	0.1189
1654412	44917	60201617	C/T	0.371	0.359	0.6627
1671151	45215	60201915	G/A	0.771	0.724	0.0765

dbSNP rs#	Position in Figure 1	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
1671152	45666	60202366	T/G	0.779	0.720	0.0234
1654413	45680	60202380	T/A	0.186	0.257	0.0048
2304167	46402	60203102	C/T	0.725	0.675	0.0720
1671153	46510	60203210	G/T	0.785	0.700	0.0016
2019599	46554	60203254	C/T	0.051	0.041	0.4200
1654485	46823	60203523	A/C	0.516	0.525	0.7872
1671188	47714	60204414	T/G	0.567	0.595	0.3451
1671191	48963	60205663	T/C	0.131	0.159	0.1931
1654415	49157	60205857	C/T	0.765	0.714	0.0536
1671192	49254	60205954	G/A	0.136	0.180	0.0436
2304168	49257	60205957	A/G	0.200	0.174	0.2713
1654416	49356	60206056	T/C	0.130	0.174	0.0432
1654419	55202	60211902	A/G	0.672	0.627	0.1243
1654420	55527	60212227	T/A	0.704	0.669	0.2206
1613662	55916	60212616	G/A	0.778	0.739	0.1347
2886415	56402	60213102	T/G	0.624	0.561	0.0324
2365593	56413	60213113	C/T	0.131	0.151	0.3369
2886414	56685	60213385	G/A	0.796	0.755	0.0976
1654421	56783	60213483	A/G	0.651	0.616	0.2323
1654424	58044	60214744	A/G	0.056	0.051	0.7154
1654425	58301	60215001	T/C	0.582	0.523	0.0464
892089	58382	60215082	A/G	0.654	0.622	0.2732
892090	58393	60215093	G/T	0.203	0.238	0.1618
1671196	58869	60215569	C/T	0.752	0.699	0.0518
1671198	59155	60215855	T/C	0.037	0.048	0.3599
1671199	59189	60215889	G/A	0.181	0.203	0.3495
1625609	62546	60219246	C/T	0.063	0.063	0.9551
1625689	62568	60219268	G/A	Not Allelotyped		
1654438	70983	60227683	A/G	0.885	0.890	0.7775
2569513	71465	60228165	G/A	0.777	0.723	0.0373
2569514	71538	60228238	G/A	0.409	0.340	0.0171
1671214	72144	60228844	G/A	0.345	0.392	0.1048
1671215	72340	60229040	C/A	0.596	0.541	0.0649
1054796	72527	60229227	C/G	0.198	0.121	0.0008
1654439	72968	60229668	T/G	0.757	0.712	0.0906
1671216	73397	60230097	A/G	0.239	0.273	0.1934
1671217	73553	60230253	G/A	0.166	0.217	0.0333
1671218	73720	60230420	C/T	0.508	0.450	0.0542
1654441	74190	60230890	C/T	0.893	0.901	0.6470
1654442	74687	60231387	T/G	0.955	0.947	0.5733
1671219	74699	60231399	G/A	0.232	0.217	0.5745
10666	75580	60232280	C/T	0.871	0.879	0.6976
1626971	76345	60233045	T/C	0.099	0.141	0.0329
1671221	76506	60233206	G/A	0.007	0.003	0.3854
754235	77554	60234254	G/A	0.445	0.486	0.1719
775821	77889	60234589	C/T	0.295	0.310	0.5716
3745912	77919	60234619	A/G	0.730	0.736	0.8182
775822	78866	60235566	A/G	0.983	0.991	0.2487
1059211	79061	60235761	C/T	0.231	0.205	0.2919
2124090	83777	60240477	A/C	0.681	0.772	0.0010
1671171	84360	60241060	T/G	0.222	0.259	0.1467
1671170	84631	60241331	T/A	0.546	0.517	0.3441
1654444	85775	60242475	G/T	0.170	0.208	0.1062
2365721	87153	60243853	G/A	0.443	0.409	0.2552
1654446	89650	60246350	A/G	0.430	0.388	0.1582
1654447	89895	60246595	A/G	0.647	0.624	0.4372
1671176	90103	60246803	C/A	0.246	0.274	0.2919
1654448	90234	60246934	G/A	0.136	0.164	0.1947

dbSNP rs#	Position in Figure 1	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
1671178	90309	60247009	G/A	0.037	0.036	0.9262
1654449	90376	60247076	G/A	0.167	0.186	0.4301
1671182	90925	60247625	C/T	0.211	0.207	0.8737
1654451	91561	60248261	A/T	0.734	0.691	0.1110
1654452	91605	60248305	G/A	0.322	0.351	0.3085
1671169	92954	60249654	T/C	0.532	0.554	0.4612
1654459	94228	60250928	A/G	0.137	0.158	0.3262
269909	Not Mapped		G/C	0.777	0.799	0.3619
269910	Not Mapped		T/G	0.234	0.190	0.0702
776251	Not Mapped		G/A	0.013	0.008	0.3916
892088	Not Mapped		A/G	0.540	0.523	0.5742
892091	Not Mapped		C/G	0.471	0.478	0.8145
1043680	Not Mapped		C/G	0.544	0.561	0.5849
1064675	Not Mapped		A/G	0.193	0.210	0.4668
1671187	Not Mapped		T/A	0.607	0.594	0.6661
2116883	Not Mapped		T/C	0.177	0.227	0.0409
2163833	Not Mapped		G/A	0.156	0.209	0.0227

[0255] Figure 14 shows the proximal SNPs in and around the GP6 gene. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 14 can be determined by consulting Table 10. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0256] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than 10^{-8} were truncated at that value.

[0257] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link (http address: www.ncbi.nlm.nih.gov/LocusLink/) are provided on the graph. The exons and

introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is place at the 3' end of each gene to show the direction of transcription.

Additional Genotyping

[0258] In addition to the SNP rs1671152, another SNP (rs1654416) was genotyped in the discovery cohort and found to be associated with breast cancer with a p-value of 0.0737. See Table 13.

[0259] The methods used to verify and genotype the proximal SNP of Table 13 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 11 and Table 12, respectively.

TABLE 11

dbSNP rs#	First PCR primer	Second PCR primer
1654416	ACGTTGGATGTCTGCTGAGCATGAAATGCC	ACGTTGGATGCTGAACTGACCGTCTCATTC

TABLE 12

dbSNP rs#	Extend Primer	Term Mix
1654416	TGACCGTCTCATTCACAAAC	ACT

[0260] Table 13, below, shows the case and control allele frequencies along with the p-values for the SNPs genotyped. The disease associated allele of column 4 is in bold and the disease associated amino acid of column 5 is also in bold. The chromosome position provided corresponds to NCBI's Build 33.

TABLE 13: Genotpying Results

dbSNP rs#	Position in Figure 1	Chromo-some Position	Alleles (A1/A2)	Amino Acid Change	AF F case	AF F control	p-value	Odds Ratio
1654416	49356	60206056	T/C	E237K	T = 0.840 C = 0.160	T = 0.800 C = 0.200	0.0737	0.75

Example 4
LAMA4 Proximal SNPs

[0261] It has been discovered that a polymorphic variation (FCH-1159) in a region that encodes laminin, alpha 4 (LAMA4) is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (FCH-1159) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. Approximately sixty-eight allelic variants located within the LAMA4 region were identified and allelotyped. The polymorphic variants are set forth in Table 14. The chromosome position provided in column four of Table 14 is based on Genome "Build 33" of NCBI's GenBank.

TABLE 14

dbSNP rs#	Chromosome	Position in Figure 2	Chromosome Position	Allele Variants
969138	6	184	112446684	C/G
1418499	6	506	112447006	T/C
2157550	6	3981	112450481	C/G
764587	6	7815	112454315	A/G
3734287	6	7875	112454375	G/A
2032565	6	10775	112457275	T/A
2032566	6	10786	112457286	T/C
1050349	6	11013	112457513	G/C
LAMA4_SNP1	6	11020	112457520	C/T
2032568	6	11101	112457601	A/G
2072019	6	14171	112460671	A/G
2072020	6	14278	112460778	G/A
1480646	6	16512	112463012	G/A
2072026	6	16706	112463206	C/T
763247	6	18442	112464942	A/C
744006	6	20286	112466786	C/T
3822941	6	21591	112468091	G/A
3777942	6	22275	112468775	A/T
3734286	6	25318	112471818	C/G
3777941	6	27997	112474497	C/T
2277084	6	29840	112476340	A/G
3798359	6	31088	112477588	C/G
3798357	6	31258	112477758	T/C
2227237	6	32367	112478867	G/A
2213838	6	32427	112478927	T/C
3752577	6	33671	112480171	G/A
LAMA4_SNP2	6	38796	112485296	C/T
971402	6	41530	112488030	A/G
971405	6	41874	112488374	A/T
2051649	6	44161	112490661	A/G
1050348	6	47502	112494002	C/T

dbSNP rs#	Chromosome	Position in Figure 2	Chromosome Position	Allele Variants
2157544	6	51089	112497589	T/G
2157545	6	51205	112497705	C/T
2213839	6	53645	112500145	G/A
LAMA4_SNP4	6	54280	112500780	G/T
3777934	6	57610	112504110	A/G
1894681	6	57740	112504240	A/G
2301512	6	60812	112507312	G/C
2301513	6	60837	112507337	T/C
2072029	6	64448	112510948	T/C
764071	6	65249	112511749	T/G
2269646	6	65482	112511982	T/C
LAMA4_SNP5	6	66535	112513035	C/T
2072022	6	66789	112513289	T/C
LAMA4_SNP6	6	67214	112513714	C/A
2237238	6	68347	112514847	C/A
3777932	6	69060	112515560	T/C
3777929	6	70100	112516600	A/G
3777928	6	70215	112516715	G/T
2157546	6	73687	112520187	C/G
3948760	6	73732	112520232	A/G
2237241	6	74183	112520683	T/C
2237242	6	74813	112521313	A/G
2237244	6	78136	112524636	T/C
3777927	6	79540	112526040	C/T
3777926	6	79655	112526155	A/G
3777925	6	79731	112526231	T/C
2239849	6	82111	112528611	G/A
2239850	6	82155	112528655	T/G
2237247	6	83479	112529979	G/A
2282853	6	84511	112531011	G/C
2282854	6	85290	112531790	T/C
2213840	6	90620	112537120	G/A
2068770	6	91127	112537627	A/G
2237248	6	92095	112538595	G/A
2237249	6	92679	112539179	A/G
2345808	6	94839	112541339	T/G
2157547	6	95220	112541720	G/C

Assay for Verifying and Allelotyping SNPs

[0262] The methods used to verify and allelotype the proximal SNPs of Table 14 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 15 and Table 16, respectively. The methods used to verify and allelotype the

proximal SNPs of Table 14 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 15 and Table 16, respectively.

TABLE 15

dbSNP rs#	Forward PCR primer	Reverse PCR primer
LAMA4_SNP5	ACGTTGGATGACAGTTCTTGGCTATCCTGG	ACGTTGGATGACTGGCCAGTGTAGGAATTG
LAMA4_SNP6	ACGTTGGATGGAAAGGGATTGACTCAGGAG	ACGTTGGATGCTTCCTTCACCTGAAGATGG
LAMA4_SNP4	ACGTTGGATGTTGAAGGACTGATCTATGGG	ACGTTGGATGAAAGCAACAGACAAGGCAAG
LAMA4_SNP1	ACGTTGGATGCAGACTGGAAATGCGCAATG	ACGTTGGATGCGTATCTTCAAGATGCACAG
1050348	ACGTTGGATGTGTTCATGTCTTCGGCATCC	ACGTTGGATGCAGCTGGATGACTACAATGC
LAMA4_SNP2	ACGTTGGATGAGGAATGCTTACAACGGAGG	ACGTTGGATGAACTCCCTTCATCCTTCCTC
744006	ACGTTGGATGTTGCCTTGAAGGTAGGCATG	ACGTTGGATGGGGTTAGCAGCTTAACTTTC
763247	ACGTTGGATGCCGGCCAAGACCAATACATC	ACGTTGGATGTGCAGACATGCACTATTCTC
764071	ACGTTGGATGCCACTTGGAAAGATTCAAGG	ACGTTGGATGTATTGTGACTTCTGCAGAAC
764587	ACGTTGGATGCAACATAGACCAGAAGTGGG	ACGTTGGATGTTCACATACGGAAGGCCTTG
969138	ACGTTGGATGACTGGACCAAGGTAGATCAC	ACGTTGGATGCTCAGGCTAATCTCTCTAGG
971402	ACGTTGGATGCCACTTTTCTGTGGAAATATC	ACGTTGGATGCAAGTTAATGAGTTTCTCCC
971405	ACGTTGGATGAAACAGTGCTTTTGAAGGAG	ACGTTGGATGCTATCTCCAAAGGGTAACAG
1050348	ACGTTGGATGCAGCTGGATGACTACAATGC	ACGTTGGATGTGTTCATGTCTTCGGCATCC
1050349	ACGTTGGATGCTATGATTTTGGATTCAGCG	ACGTTGGATGACCTCATGGTATTTTGCATC
1158747	ACGTTGGATGTTGAAGGACTGATCTATGGG	ACGTTGGATGAAAGCAACAGACAAGGCAAG
1418499	ACGTTGGATGACCATAGGGAACTAGAAATC	ACGTTGGATGCTTTAAGATAGATTCCCAGGG
1480646	ACGTTGGATGCAGTGTCTCTTCCTTTCCAG	ACGTTGGATGCAAATTTCCACGAGCCTGAG
1894681	ACGTTGGATGTGGGATTCCCCTAAAGGATG	ACGTTGGATGAAGATCAGCAGCACCAAAGG
2032565	ACGTTGGATGAAAGAGCAACTGAAGGACCC	ACGTTGGATGTAAATTGGAACATCAACAGG
2032566	ACGTTGGATGTAAATTGGAACATCAACAGG	ACGTTGGATGAAAGAGCAACTGAAGGACCC
2032567	ACGTTGGATGCGTATCTTCAAGATGCACAG	ACGTTGGATGAGACTGGAAATGCGCAATGG
2032568	ACGTTGGATGACTCGCATAACAGATGTTCC	ACGTTGGATGTAACCATTGCGCATTTCCAG
2051649	ACGTTGGATGACCTGCTGAAAACCAACACC	ACGTTGGATGGGAGAGGAGAACCCTGGAC
2068770	ACGTTGGATGCACTTCACGTACTTCACTGG	ACGTTGGATGAGTTTGCTCCTATGTGGCTC
2072019	ACGTTGGATGAGGTCCACAGAAGATGTTAG	ACGTTGGATGCACAACGGTCATTTGAACAC
2072020	ACGTTGGATGAAGTCCTGTTGTCTGCAAGG	ACGTTGGATGCAGTTGTCTTAGCACACAGG
2072022	ACGTTGGATGCAAAGAAGAAAGATGTAGTGG	ACGTTGGATGCGAAATCTGGTCCTATGAAG
2072026	ACGTTGGATGTCCTATCACCATCACACTAC	ACGTTGGATGCAGCATCAAACAGAATAGGC
2072029	ACGTTGGATGTCCTTGCAGACTGATACTCC	ACGTTGGATGTCACTCACTCCTTGCTAAGC
2157544	ACGTTGGATGCATATGTAGTAGGAATGAGGG	ACGTTGGATGTGAGGCTCAAAGGGATTAGG
2157545	ACGTTGGATGTCTGGTCAACCACATAGATC	ACGTTGGATGTGTTCTACTGCAGCTCCAAG
2157546	ACGTTGGATGTCCACTTGTACAGAATGGAG	ACGTTGGATGCATTTACTCAGTGCCAGGTC
2157547	ACGTTGGATGCCATACCATTTACTTCTGCC	ACGTTGGATGAGGCAAGTACACATACAATG
2157550	ACGTTGGATGCACACACACATTTTAATTGCC	ACGTTGGATGTTGTTCAGAATTACATGATG
2213838	ACGTTGGATGAAAGGACTTGAGGGTGATTG	ACGTTGGATGGCAACAAACAGTGTTCCAGC
2213839	ACGTTGGATGAGTCACAGTTCAGTCCCAAC	ACGTTGGATGGGGCAATTTTCTAGTCCAGC
2213840	ACGTTGGATGCTTTCGCACAAGGCTCTATC	ACGTTGGATGAAGTCTGTGTTTAAGCCCCC
2227237	ACGTTGGATGGATGTCTCTAAGTTGAAATGC	ACGTTGGATGATATCAATCACCCTCAAGTC
2237238	ACGTTGGATGCAGAGGCTGAAGGAACATAC	ACGTTGGATGTCTGTAATCCCAGGACCCTA
2237241	ACGTTGGATGTCAGCAGGGCTCTATCTAAG	ACGTTGGATGCCAAGCAGTATTGCTAATGG

dbSNP rs#	Forward PCR primer	Reverse PCR primer
2237242	ACGTTGGATGCCTCACCATTGTGTTTAGGC	ACGTTGGATGTGACTATTTCCGCTTGGCTC
2237244	ACGTTGGATGGAGAAAAATAGACTCGGCCC	ACGTTGGATGCACAGACGCAGGATTTGGAT
2237247	ACGTTGGATGCTGCTTCTCCAGTAATGTTG	ACGTTGGATGGTGTTAGTAACACTGATGCC
2237248	ACGTTGGATGCCCTCCCCAGATATCATTAG	ACGTTGGATGCATATCCACAGCCTAATCAC
2237249	ACGTTGGATGAAATGCTTCCTACTGCAATC	ACGTTGGATGTGGAGAGTTGTGGTTGATGG
2239849	ACGTTGGATGGACATCAGATCAGACAGCAC	ACGTTGGATGACTTTCTGGCATTGACTGGG
2239850	ACGTTGGATGGCCCAGGAAAAATTAATTCAC	ACGTTGGATGGCAGTACGGATTAGCATGAG
2269646	ACGTTGGATGTCACCTCACTTTTGAAGAGC	ACGTTGGATGTCTGGTTAGGCTTCAGTTAG
2277084	ACGTTGGATGGAGGGTAAAAATGACAGCAG	ACGTTGGATGTTTTGCTTGGTGTTTAGCAG
2282853	ACGTTGGATGTCTTGACCTTCCTGGTTTTC	ACGTTGGATGTATCAGAGCTAGAAGAAACC
2282854	ACGTTGGATGTAGCCAGTGGTTAAGAAAGC	ACGTTGGATGTTCTCATGTTGGGGAGACAC
2301512	ACGTTGGATGATCTGAGTGGTTTCAGGAGG	ACGTTGGATGACCTGTTGGAACACATGAAG
2301513	ACGTTGGATGCTGGCGGGTAGTGTCTTCAT	ACGTTGGATGCTTTGAAATTGTTCTTGTCC
2345808	ACGTTGGATGTTCTGGGATTTAAAGGAGGC	ACGTTGGATGCCAAACATTTCTTGTTGGAC
3734286	ACGTTGGATGACCTTACACTCCAGTGAATC	ACGTTGGATGGCCGTTAAGCAACTACAAGC
3734287	ACGTTGGATGCAGTGGAGAAGATGAAACCC	ACGTTGGATGCCCACTTCTGGTCTATGTTG
3752577	ACGTTGGATGCATGGCTGAGGTTACTTAGG	ACGTTGGATGGAATGCGTCAGGGATTTATG
3777925	ACGTTGGATGCTACAAGTCTAACAGTCAGAG	ACGTTGGATGTTACAGAGCAAGGTCTGAGG
3777926	ACGTTGGATGGTGAGTACCATCCCTTTTGC	ACGTTGGATGCTGTTAAACTGCCTCAGACC
3777927	ACGTTGGATGAAACGAATGCTTGAGAGCAG	ACGTTGGATGGTCCTGATTTATGAGCTCCC
3777928	ACGTTGGATGTTCACACGTAGACCCTGTTG	ACGTTGGATGTCAGGAGTTGAGCAAGCTAG
3777929	ACGTTGGATGGCTGTCTTTGGGATTAAAT	ACGTTGGATGTTCATAAAGAAGTGGAGAGC
3777932	ACGTTGGATGTCCCAGACCTTAAGATTCCC	ACGTTGGATGTATTAGGCTCTTTGGCCGAC
3777934	ACGTTGGATGCAAGATCCAGATGGTGAGGG	ACGTTGGATGCAAGGTCAGAGTGTCACTGG
3777941	ACGTTGGATGGCTTCCTGAGATTATATTGAC	ACGTTGGATGCTCCATTCCAAATTCCTTTC
3777942	ACGTTGGATGTCATGACAAATCATGACTAG	ACGTTGGATGTCAGATACAAGTGAAGGTAG
3798357	ACGTTGGATGTCCCAATTCAGGAAATGGTG	ACGTTGGATGTGCTTGGTATACCATGCCTG
3798359	ACGTTGGATGTTCCTCAGCACACAGCCCCA	ACGTTGGATGATGAACCTTACACAGGCCAG
3822941	ACGTTGGATGTATAATAAACTGATAGTTGC	ACGTTGGATGCTCTGTACTTAGGACACACG
3948760	ACGTTGGATGTCCACTTGTACAGAATGGAG	ACGTTGGATGTCCCACACTCAAAACTTTGC

TABLE 16

dbSNP rs#	Extend Primer	Term Mix
LAMA4_SNP5	ATTGCTTACGCAACACCAC	ACG
LAMA4_SNP6	AGATGGAGAGAATGCCAC	CGT
LAMA4_SNP4	GCAAGTGGGCATTCGACCA	CGT
LAMA4_SNP1	CTTCAAGATGCACAGGGCCAC	ACG
1050348	CACTTGACCAGGCCCTTAAC	ACG
LAMA4_SNP2	GGCCCGCCTGCATCTGTG	ACG
744006	CTTTCTCTCTTTCCAGG	ACG
763247	CTTTAATCCCCCACACT	ACT
764071	AGAACATATATGTTGCATTTTTTT	ACT

764587	GAAGGCCTTGCCTGTTA	ACT
969138	AGGAAGAGAATCTGATAGCC	ACT
971402	AGTTTCTCCCACTTACC	ACT
971405	AGGGTAACAGAATGATTAAAA	CGT
1050348	TTCGGCATCCCTGACAT	ACT
1050349	CGTATCTTCAAGATGCACA	ACT
1158747	GCAAGTGGGCATTCGACCA	CGT
1418499	GGGCAGAATTACTGAATCAAG	ACT
1480646	CAGCAGACTCTGATGTGGC	ACG
1894681	GGGAGCATCTTTTGAGC	ACT
2032565	CAACAGGAAAAATACATCCA	CGT
2032566	CAACCCTAGGAAAACATTT	ACT
2032567	TTCTATGATTTTGGATTCAGC	ACT
2032568	ACATACTCTGAGGAGAGAAAG	ACT
2051649	GAACCCTGGACAAGAAT	ACT
2068770	ATGTGGCTCAAACATCCGAA	ACT
2072019	TTTGAACACTACAGTTTCTGTTAT	ACT
2072020	AAACAATCCATTTAACATACCTA	ACG
2072022	GCAAATGAATTCTGGGA	ACT
2072026	TGAAAGTCTTTGAGGTGTT	ACG
2072029	CCTGGCAATGATCAACCCCC	ACT
2157544	CTAAATATTAGCAGACTGAAATAC	ACT
2157545	GCTGGCATAAATGAAATTG	ACG
2157546	GTGCCAGGTCCCACACT	ACT
2157547	GTACACATACAATGATTTTACTC	ACT
2157550	TTACATGATGAATATTATGGAAGT	ACT
2213838	TTCCAGCATGATTCTAAGACA	ACT
2213839	CAACTTGAGATACAGTAAAAATT	ACG
2213840	TGAAATGAATTCTCCAATAGAC	ACG
2227237	ACCCTCAAGTCCTTTTG	ACG
2237238	TCCCAGGACCCTAAAAAAGT	CGT
2237241	CAGTATTGCTAATGGGTGTTC	ACT
2237242	TGTCTCTAGGGCACTACATATC	ACT
2237244	GAAATAATGCTTCAGGGG	ACT
2237247	ATGCCTTCTAATGCATTCATTTTA	ACG
2237248	CCTAATCACATAAACCAGGAA	ACG
2237249	GAAAACAAGAGAGGGAAG	ACT
2239849	TGTGACTCCTCATGCTAATC	ACG
2239850	CCAGTCAATGCCAGAAA	ACT
2269646	CAGTTAGACTGAAACGCACA	ACT
2277084	TGTGTCATTTAAATCCTTCA	ACT
2282853	GAGCTAGAAGAAACCTGAAAG	ACT
2282854	GTTGGTGTCCAAATGGCA	ACT
2301512	ACATGAAGACACTACCC	ACT
2301513	TGTTCTTGTCCAAAATTACCT	ACT
2345808	GACATTTAGGTTATTTCCAAATTT	ACT

3734286	CATCAGAGAGAATTGAAGT	ACT
3734287	GGAATTCAGGCATACAC	ACG
3752577	AGAAATAGATGGAGCCAAAAG	ACG
3777925	GGATGGGACTGAAACTC	ACT
3777926	CTCTGTAATTTTTCATGTATGATA	ACT
3777927	ATGAGCTCCCTTCACTC	ACG
3777928	TGAGCAAGCTAGAGAGTA	CGT
3777929	AAGTGGAGAGCATTTACAT	ACT
3777932	TTTGGCCGACTGAAATG	ACT
3777934	GGTCAGAGTGTCACTGGGCTACA	ACT
3777941	TCCCAAATTTCCTTTTCA	ACG
3777942	ACAAGTGAAGGTAGTATTGT	CGT
3798357	GCCTGGCATCTGCTAATC	ACT
3798359	GCAAAGGCAGAGACTAT	ACT
3822941	ACACACGATGTTTCTCCAG	ACG
3948760	ACAGTTTTATGAGACAGGTA	ACT

Genetic Analysis of Allelotyping Results

[0263] Allelotyping results are shown for cases and controls in Table 17. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where "AF" is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP rs969138 has the following case and control allele frequencies: case A1 (C) = 0.893; case A2 (G) = 0.107; control A1 (C) = 0.866; and control A2 (G) = 0.134, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped.

TABLE 17

dbSNP rs#	Position in Figure 2	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
969138	184	112446684	C/G	0.107	0.134	0.1800
1418499	506	112447006	T/C	0.474	0.545	0.0181
2157550	3981	112450481	C/G	0.238	0.294	0.0356
764587	7815	112454315	A/G	0.263	0.192	0.0054
3734287	7875	112454375	G/A	0.645	0.729	0.0032
2032565	10775	112457275	T/A	0.271	0.303	0.2393
2032566	10786	112457286	T/C	0.560	0.527	0.2791
1050349	11013	112457513	G/C	0.789	0.806	0.4918
LAMA4_SNP1	11020	112457520	C/T	0.556	0.480	0.0109
2032568	11101	112457601	A/G	0.280	0.336	0.0429
2072019	14171	112460671	A/G	0.658	0.608	0.0830
2072020	14278	112460778	G/A	0.171	0.170	0.9646
1480646	16512	112463012	G/A	0.329	0.329	0.9962
2072026	16706	112463206	C/T	0.326	0.367	0.1525
763247	18442	112464942	A/C	0.293	0.360	0.0168

dbSNP rs#	Position in Figure 2	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
744006	20286	112466786	C/T	0.697	0.670	0.3444
3822941	21591	112468091	G/A	0.694	0.670	0.3923
3777942	22275	112468775	A/T	0.234	0.254	0.4484
3734286	25318	112471818	C/G	0.240	0.249	0.7505
3777941	27997	112474497	C/T	0.590	0.573	0.5688
2277084	29840	112476340	A/G	0.170	0.193	0.3128
3798359	31088	112477588	C/G	0.189	0.194	0.8418
3798357	31258	112477758	T/C	0.607	0.589	0.5393
2227237	32367	112478867	G/A	0.699	0.678	0.4600
2213838	32427	112478927	T/C	0.619	0.606	0.6603
3752577	33671	112480171	G/A	0.866	0.856	0.6307
LAMA4_SNP2	38796	112485296	C/T	0.480	0.441	0.1869
971402	41530	112488030	A/G	0.406	0.397	0.7632
971405	41874	112488374	A/T	0.396	0.370	0.3662
2051649	44161	112490661	A/G	0.620	0.596	0.4216
1050348	47502	112494002	C/T	0.437	0.518	0.0071
2157544	51089	112497589	T/G	0.344	0.396	0.0695
2157545	51205	112497705	C/T	0.143	0.116	0.1789
2213839	53645	112500145	G/A	0.329	0.387	0.0415
LAMA4_SNP4	54280	112500780	G/T	0.339	0.392	0.0645
3777934	57610	112504110	A/G	0.065	0.068	0.8646
1894681	57740	112504240	A/G	0.433	0.382	0.0840
2301512	60812	112507312	G/C	0.089	0.092	0.8857
2301513	60837	112507337	T/C	0.688	0.741	0.0514
2072029	64448	112510948	T/C	0.765	0.782	0.5146
764071	65249	112511749	T/G	0.431	0.502	0.0175
2269646	65482	112511982	T/C	0.438	0.379	0.0458
LAMA4_SNP5	66535	112513035	C/T	0.099	0.071	0.1061
2072022	66789	112513289	T/C	0.956	0.954	0.9004
LAMA4_SNP6	67214	112513714	C/A	0.410	0.452	0.1649
2237238	68347	112514847	C/A	0.111	0.099	0.5343
3777932	69060	112515560	T/C	0.080	0.073	0.6848
3777929	70100	112516600	A/G	0.311	0.364	0.0656
3777928	70215	112516715	G/T	0.414	0.364	0.0861
2157546	73687	112520187	C/G	0.331	0.415	0.0045
3948760	73732	112520232	A/G	0.187	0.269	0.0015
2237241	74183	112520683	T/C	0.384	0.447	0.0360
2237242	74813	112521313	A/G	0.434	0.391	0.1478
2237244	78136	112524636	T/C	0.492	0.515	0.4374
3777927	79540	112526040	C/T	0.110	0.042	0.0001
3777926	79655	112526155	A/G	0.854	0.858	0.8519
3777925	79731	112526231	T/C	0.774	0.771	0.9216
2239849	82111	112528611	G/A	0.734	0.771	0.1596
2239850	82155	112528655	T/G	0.851	0.936	0.0000
2237247	83479	112529979	G/A	0.097	0.093	0.8578
2282853	84511	112531011	G/C	0.775	0.783	0.7522
2282854	85290	112531790	T/C	0.078	0.014	0.0000
2213840	90620	112537120	G/A	0.738	0.806	0.0077
2068770	91127	112537627	A/G	0.332	0.277	0.0479
2237248	92095	112538595	G/A	0.212	0.153	0.0111
2237249	92679	112539179	A/G	0.339	0.270	0.0128
2345808	94839	112541339	T/G	0.182	0.114	0.0019
2157547	95220	112541720	G/C	0.176	0.128	0.0258

[0264] Figure 15 shows the proximal SNPs in and around the LAMA4 region. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 15 can be determined by consulting Table 17. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0265] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than 10^{-8} were truncated at that value.

[0266] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link (http address: www.ncbi.nlm.nih.gov/LocusLink/) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is place at the 3' end of each gene to show the direction of transcription.

Example 5
CHGB/C20orf154 Proximal SNPs

[0267] It has been discovered that a polymorphic variation (rs454422) in a gene region that includes CHGB and C20orf154 is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs454422) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. A total of ninety-eight allelic variants located within or nearby the CHGB/C20orf154 region were identified and allelotyped.

The polymorphic variants are set forth in Table 18. The chromosome position provided in column four of Table 18 is based on Genome "Build 33" of NCBI's GenBank.

TABLE 18

dbSNP rs#	Chromosome	Position in Figure 3	Chromosome Position	Allele Variants
2300427	20	186	5842586	T/G
571039	20	1332	5843732	G/T
236143	20	1893	5844293	T/C
236145	20	2786	5845186	C/G
236146	20	2962	5845362	A/G
446658	20	3377	5845777	A/C
500277	20	5522	5847922	A/G
2268339	20	5621	5848021	A/G
454328	20	5889	5848289	G/A
236148	20	7531	5849931	T/C
236149	20	8268	5850668	T/C
910122	20	8923	5851323	G/A
881118	20	8988	5851388	A/C
236151	20	9117	5851517	G/A
236152	20	9448	5851848	C/G
CHGB_SNP1	20	9494	5851894	G/A
742710	20	9628	5852028	G/A
742711	20	9640	5852040	T/C
236154	20	11072	5853472	T/G
236155	20	11150	5853550	A/G
54144	20	11379	5853779	C/A
540717	20	11692	5854092	G/A
236158	20	12056	5854456	T/C
236159	20	12104	5854504	G/A
446614	20	14160	5856560	T/C
1343180	20	14836	5857236	A/G
1039542	20	14980	5857380	A/C
400735	20	15165	5857565	A/G
1039543	20	15315	5857715	A/G
440005	20	15624	5858024	A/C
394604	20	15796	5858196	C/T
546106	20	15939	5858339	T/C
452749	20	16581	5858981	C/T
364652	20	17045	5859445	T/C
403727	20	18501	5860901	A/G
236160	20	21800	5864200	A/G
236161	20	21966	5864366	A/G
236162	20	22134	5864534	A/G
236163	20	22181	5864581	A/G
236164	20	23028	5865428	A/G
236165	20	23312	5865712	A/G

dbSNP rs#	Chromosome	Position in Figure 3	Chromosome Position	Allele Variants
236166	20	23573	5865973	T/A
236167	20	23858	5866258	A/G
183535	20	23888	5866288	G/A
236168	20	23990	5866390	T/C
236169	20	24073	5866473	A/G
236171	20	25330	5867730	G/A
236173	20	26473	5868873	T/C
236175	20	27958	5870358	C/T
236176	20	28421	5870821	A/G
451571	20	28804	5871204	C/T
236177	20	29322	5871722	C/A
1005517	20	30819	5873219	G/C
236179	20	31956	5874356	G/A
236180	20	32592	5874992	G/A
236181	20	32818	5875218	G/C
236182	20	32880	5875280	G/T
236183	20	33244	5875644	G/C
236184	20	33845	5876245	A/G
236185	20	34272	5876672	G/A
236187	20	34931	5877331	T/C
1394095	20	36870	5879270	T/G
236189	20	37790	5880190	T/C
236110	20	38708	5881108	T/G
236111	20	39135	5881535	T/C
236112	20	39919	5882319	A/G
236113	20	40166	5882566	C/T
236114	20	40985	5883385	A/G
236115	20	41049	5883449	A/G
236116	20	41935	5884335	C/T
236117	20	42775	5885175	A/C
236118	20	43807	5886207	T/C
236119	20	44254	5886654	A/G
3761873	20	44814	5887214	A/C
236120	20	45249	5887649	T/G
451417	20	47599	5889999	C/A
379418	20	47807	5890207	G/A
2326680	20	48555	5890955	A/C
409035	20	49249	5891649	G/A
454422	20	49293	5891693	A/C
236102	20	57566	5899966	A/C
236103	20	63587	5905987	T/C
236104	20	64560	5906960	C/T
236105	20	65432	5907832	C/G
236106	20	66291	5908691	T/C
236107	20	71331	5913731	A/T
180477	20	73344	5915744	A/T
236108	20	74159	5916559	C/T

dbSNP rs#	Chromosome	Position in Figure 3	Chromosome Position	Allele Variants
236109	20	74564	5916964	T/C
236121	20	78194	5920594	A/G
236122	20	79128	5921528	T/C
236123	20	79393	5921793	C/T
236124	20	81579	5923979	G/A
236125	20	82574	5924974	C/T
2876003	20	85309	5927709	G/C
CHGB_SNP2	20	87076	5929476	A/G
2423131	20	87844	5930244	C/A
2206817	20	90241	5932641	T/C

Assay for Verifying and Allelotyping SNPs

[0268] The methods used to verify and allelotype the sixty-three proximal SNPs of Table 18 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 19 and Table 20, respectively.

TABLE 19

dbSNP rs#	Forward PCR primer	Reverse PCR primer
54144	ACGTTGGATGAGCAGCCATCACATGATCTG	ACGTTGGATGCCAATTGTGTCAAACATTTAATGAA
183535	ACGTTGGATGTGGTCTAAGTCCTGCACAAG	ACGTTGGATGGACAGAAAGAGTGTGCAGTC
236104	ACGTTGGATGCTCCCTGAACTTCCCATTTC	ACGTTGGATGCAGGGAGCTTACTACAAAGC
236105	ACGTTGGATGGGTAGCCACATTTGACACAG	ACGTTGGATGAAGTGGCCTGGAAACCAATG
236113	ACGTTGGATGGCTAATACATCCTAAAGGAAC	ACGTTGGATGCTTTATTGGAAATCTGTTCG
236119	ACGTTGGATGTTTCTTCTTGTCTCACAGGC	ACGTTGGATGCTGTAGATTTTCCTTTTGGC
236120	ACGTTGGATGCAGGATAATGATCATCCCTG	ACGTTGGATGACTGGCACAATTAGTGTCTG
236122	ACGTTGGATGTAGCTATCCCATTGTTGAGG	ACGTTGGATGGCTAAAGTCTGAAAACACTAG
236143	ACGTTGGATGCCTACCCCAAATAGGTAAAG	ACGTTGGATGAGGAAGTAAGTTTTGGGAGG
236145	ACGTTGGATGAGAATGTGAGCAAGGGATGC	ACGTTGGATGAATGGCACAGCTATCCTCAG
236146	ACGTTGGATGGTAAATCTGTATCTCCGCCC	ACGTTGGATGGACTTGATTACGTGACCTGG
236149	ACGTTGGATGCCAATGATCAAACTGAAATCG	ACGTTGGATGGGACTGGATTAGATGAATTC
236160	ACGTTGGATGCACCCACAGCTATCTGAGTT	ACGTTGGATGTCCAAAAGGAGTGGGTAGAG
236162	ACGTTGGATGAATGGAGTGTTTGCATGTTG	ACGTTGGATGCATGGATTTTTAGGACATGCG
236163	ACGTTGGATGGCCAAAAATAGCCTTTTCTC	ACGTTGGATGAAACAACATGCAAACACTCC
236164	ACGTTGGATGTTGCAAGCTGGTGTCACACA	ACGTTGGATGAGACCAGCCACTACTGATTC
236166	ACGTTGGATGCTCCTGTCATAGAATAGGCC	ACGTTGGATGGCATAGCACATGCTATTTGG
236167	ACGTTGGATGATCAAAGTCTTGTGCAGGAC	ACGTTGGATGGCACTTTAGGGACATTTGAC
236176	ACGTTGGATGGCATAGATAGCTTAATCATGG	ACGTTGGATGAAACCAATAGAAGCAGGTTG
236177	ACGTTGGATGCATTTGAGGATCGGAGTGAG	ACGTTGGATGCCCTGTGTCTGCAAATTTGG
236180	ACGTTGGATGCCTAGCACTTGGGAATTAGG	ACGTTGGATGGATGCGTGAAATAGATGCTC

dbSNP rs#	Forward PCR primer	Reverse PCR primer
236182	ACGTTGGATGCTCTCTAGTTCCTTTGTTGC	ACGTTGGATGATTTCAAAAGTGGTCTCCAC
236183	ACGTTGGATGTAAAAGAGAAATCCCACAGG	ACGTTGGATGCTCAGCAGATGTTAGTTTTC
236184	ACGTTGGATGCTAGGGACCCAGCAAATAAC	ACGTTGGATGACCATCTGAGGGAAATCCTG
379418	ACGTTGGATGTCTGGCCTCTAAGCTAAAGG	ACGTTGGATGCAGTGGTGTTGATAGATGGG
400735	ACGTTGGATGACCTTCTCAGGTTCACGTTC	ACGTTGGATGAAGGGCACCATGCCATTAAC
409035	ACGTTGGATGTGCTATGGGTATGCAAGGTG	ACGTTGGATGGGTTGTTGAAGGAGCGAGAG
440005	ACGTTGGATGATGGCTCTAAAAAGCTGTCC	ACGTTGGATGCAGCCTTCTTCCTGATACAG
446614	ACGTTGGATGTAAAGCCCAGGGCTAAAGAC	ACGTTGGATGAGAAGATGGCCAGAAAGGAG
451417	ACGTTGGATGCCAGAGTCATCGTTATCACC	ACGTTGGATGCCTCACTAAGGATTCAACCC
454422	ACGTTGGATGCAGCTTTTGAGGCACTTTCC	ACGTTGGATGAGCACCTTGCATACCCATAG
500277	ACGTTGGATGAGTTTCCCTACGTCTCTCTC	ACGTTGGATGAGTAACTCTAGCCTCTGCTC
540717	ACGTTGGATGCATCCAAAACCCAACAAATCC	ACGTTGGATGAGAGAGGTGTGTGACTTTTC
546106	ACGTTGGATGTTATAGCACTGATGGGCTCC	ACGTTGGATGCTGTGACATACTTTTCCAGG
571039	ACGTTGGATGATTCCTGTAGCAGGCAACTG	ACGTTGGATGGCTAGCTCTACTCTCTTCTC
1039542	ACGTTGGATGTGAGGTTCTGTCTGAACACC	ACGTTGGATGTGGCTGCAATGGCTAACTTC
1039543	ACGTTGGATGATCTGACTCAGAAGAAGAGC	ACGTTGGATGGGCATTAATGGAGGTTATGC
1343180	ACGTTGGATGAGATGGCAACAGCAACACAG	ACGTTGGATGCCAACAGCAGCTTCACAATC
CHGB_SNP2	ACGTTGGATGAAATGGTATGTTTGTGTTCC	ACGTTGGATGTAATTTTTCCCCCCCAAATC
rs384578	ACGTTGGATGAAATGGTATGTTTGTGTTCC	ACGTTGGATGTAATTTTTCCCCCCCAAATC
rs742710	ACGTTGGATGAGAAAGTGAGGAAGAGAGGG	ACGTTGGATGATGAAATAGGCACGTGGCTC
rs742711	ACGTTGGATGATGAAATAGGCACGTGGCTC	ACGTTGGATGAGAAAGTGAGGAAGAGAGGG
rs881118	ACGTTGGATGTATAGCTGAAGCCTGCTTTC	ACGTTGGATGCAGTGAAGAGAAACACCTTG
rs910122	ACGTTGGATGAAGGTGTTTCTCTTCACTGC	ACGTTGGATGGGAGGGAGAGAACTATCAAA
CHGB_SNP1	ACGTTGGATGTCACTCTGAGGTCTTGGAGC	ACGTTGGATGTAAAGGGTTATCCAGGCGTC
180477	ACGTTGGATGGGAAGTAATTCTCTGGGCTG	ACGTTGGATGAAGTGATCCTCCCACCTCAG
236102	ACGTTGGATGCAGCCTGTTCTCTCTGAAAC	ACGTTGGATGGGATGCAAGAGGTTGTAGAG
236103	ACGTTGGATGCCTGTTTAAATCGTGGCTCC	ACGTTGGATGAAACATAAGGAAGCTGAGGC
236106	ACGTTGGATGCAAGCCTTTGCAGCTCTATC	ACGTTGGATGCCTCATAAGGGCCTTTGTAC
236107	ACGTTGGATGGAAGTTTACGTAAACTCTAG	ACGTTGGATGGTGTGTGGCTTATTGTAGAG
236108	ACGTTGGATGGTATTTACTGTTGAACCCAG	ACGTTGGATGATGTGGGTAAGTTGTGCACC
236109	ACGTTGGATGAGATTACAGGCACTAGCCAC	ACGTTGGATGTCTGGGCAACATGGTGAAAC
236110	ACGTTGGATGATCGATCCAATGTTGACTGC	ACGTTGGATGTTTCAGAACAAACCCCACAG
236111	ACGTTGGATGTTCAGGAAGCAGCAACCATC	ACGTTGGATGTATGCTGTGACCTCTCCAAC
236112	ACGTTGGATGAACGAGGTCAGGAGATCAAG	ACGTTGGATGCACGCCCGGCTAATTTTTTC
236114	ACGTTGGATGGAACCAAGGAAGTCTGACTC	ACGTTGGATGAAAGCTACCAGTCATGTGCC
236115	ACGTTGGATGATCAAAGTCCATACTGCAGG	ACGTTGGATGTATGATCGTAGGCACTGGAG
236116	ACGTTGGATGTGTTGTATTACCTGACCCTG	ACGTTGGATGAAGCAAACCACTGAGTGTCC
236117	ACGTTGGATGCAATGGTGTGATCTTGCCTC	ACGTTGGATGATTAGCCAAGTGTGGCAGTG
236118	ACGTTGGATGGGTTGAGTATCCCTAATCTG	ACGTTGGATGCTTTCAGTGTCGTGTCAGGG
236121	ACGTTGGATGCAAGCTATGTCACAGTTTAAG	ACGTTGGATGAGTCTTTGCCCTTAATGTGG
236123	ACGTTGGATGATAATAAATTTAGACTTCAC	ACGTTGGATGAAAATACTGGTGCGGCCAGA
236124	ACGTTGGATGGAATTTTGTTTGGCTCACGG	ACGTTGGATGATTGCTGCTGGAAGCTTACC
236125	ACGTTGGATGCCATGCCTGAGTTATTTGC	ACGTTGGATGATGGAGAAAGTAGATAGTAG
236148	ACGTTGGATGTAAGCCCAAGTGCTGTTGAG	ACGTTGGATGCTCAGAAGTCTGATGTGTATC
236151	ACGTTGGATGTTGGCCTTTAGACTCCTGGG	ACGTTGGATGAGAAGACACATAGCCGAGAG
236152	ACGTTGGATGAATAAAGGGTTATCCAGGCG	ACGTTGGATGTGGAGCCCTGTATTCTTCAC
236154	ACGTTGGATGTTCTGACAAGTTCCTGGCTG	ACGTTGGATGGCTGCATTAGTCAACCTACC

dbSNP rs#	Forward PCR primer	Reverse PCR primer
236155	ACGTTGGATGGGTAGGTTGACTAATGCAGC	ACGTTGGATGTGAGGTCCCGAACCAATTTC
236158	ACGTTGGATGAAACTCCTGACCTCGTGATC	ACGTTGGATGCTCCTTAAGAAGATAGAGGC
236159	ACGTTGGATGGTCTCAAACTCCTGACCTCG	ACGTTGGATGAAGAAGATAGAGGCAGCTGG
236161	ACGTTGGATGGATGTTGCCTCTAGGCTAGT	ACGTTGGATGCACCATCTGACCTGTGCTAC
236165	ACGTTGGATGAAAATTAGCCATGCGTGGTG	ACGTTGGATGTTCAAGCGGTTCTCCTGCCT
236168	ACGTTGGATGTCTATGTCTCCACTTGCATG	ACGTTGGATGACACATTTGCACACACACAC
236169	ACGTTGGATGGTGACTAGAATTTTTGTGTAC	ACGTTGGATGGTGTGTGCAAATGTGTATCC
236171	ACGTTGGATGAACCTCCCACTTTGGCTTTC	ACGTTGGATGGGTCCATTTAAAGCCTGGTG
236173	ACGTTGGATGATCAACCTGCACCACCAATC	ACGTTGGATGGCTAAGATGGAAGTTGAAGTG
236175	ACGTTGGATGTTCTCCATCACTGCATCAAG	ACGTTGGATGGTTATAGCCTGTATCGCAGC
236179	ACGTTGGATGCTAAATAACAGGTTTGACTC	ACGTTGGATGGAACATTGAGAGTATCTTAT
236181	ACGTTGGATGGTGAACATGTCTTTTCTGTAC	ACGTTGGATGGGTAGAACCACTGTTTTTCG
236185	ACGTTGGATGGGGTCACTTGAATTCAGGAG	ACGTTGGATGACTGCAACCACTGCCTCTTG
236187	ACGTTGGATGTAGTGAAACTCTGTCTCTGC	ACGTTGGATGACCTGCACCAACCTTTAACC
236189	ACGTTGGATGTGGATTTACAGAAAAACTGC	ACGTTGGATGCTGTGAGACACTAGGGATAC
364652	ACGTTGGATGTTTCTGCTGGGCTGTGATAG	ACGTTGGATGGGGAAATGCTCAGCATGTAC
394604	ACGTTGGATGTATTTTGGGATGGTGTGGGC	ACGTTGGATGGAACCAGGTCTTCCTTGATG
403727	ACGTTGGATGTCACTTGAACCCAGGAGATG	ACGTTGGATGGTTTTGAGACAGAGTTTCGC
446658	ACGTTGGATGTCACTGAGTTCAACTCCTTC	ACGTTGGATGGTTCCTGCTTTACCACTTCG
451571	ACGTTGGATGTTCTGGGTGGTTGCTCTCTG	ACGTTGGATGAAGTAATGGCACACTGGAGG
452749	ACGTTGGATGTCCTACTCCAGTATGACCTC	ACGTTGGATGGAAGTCCCAACCCCTAATAC
454328	ACGTTGGATGTGCAAACTGGTGCATCAGAG	ACGTTGGATGCCTGGTATTTTCATATCGCC
742710	ACGTTGGATGGGCACATGGATATGGTGAAG	ACGTTGGATGTGCCTCTGTGATGGTGTCCC
742711	ACGTTGGATGGGCACATGGATATGGTGAAG	ACGTTGGATGAAATAGGCACGTGGCTCCCC
881118	ACGTTGGATGTATAGCTGAAGCCTGCTTTC	ACGTTGGATGTAGCAGTGAAGAGAAACACC
910122	ACGTTGGATGGTTTCTCTTCACTGCTATCT	ACGTTGGATGACACGCCATTCTGAGAAGAG
1005517	ACGTTGGATGTACTAATGTCAGTGGTAGAG	ACGTTGGATGTGAAGACACTGGCTGAAAAC
1394095	ACGTTGGATGTTCAGTGATCCAACTTCCGC	ACGTTGGATGCCAAACTCCTTGATTGGC
2206817	ACGTTGGATGGCAGAAACCCAGTGAAGTAG	ACGTTGGATGAAACCAGTTACTAACTGTAG
2268339	ACGTTGGATGATCCTGGAGATGTTATACCC	ACGTTGGATGCCTGGTGTTTAAGGCTCAAC
2300427	ACGTTGGATGAGATTACAGGCATGAGCCAC	ACGTTGGATGAAGTTAAATAAGCTCTTCTG
2326680	ACGTTGGATGAGGCTAATTCCTTCTCCTGG	ACGTTGGATGTCGTGCAACATCACTGTGTC
2423131	ACGTTGGATGATGCCTGCCTTACGAGAATG	ACGTTGGATGTGTCACTAGAATATGTGAAC
2876003	ACGTTGGATGGCAAAGACTAAGAGTCTGTAG	ACGTTGGATGCTGAGCCAGATTCTGACATT
3761873	ACGTTGGATGCTGTCCCTCTTAGAGCAATG	ACGTTGGATGCTATGAGCCTTTGACACAGC

TABLE 20

dbSNP rs#	Extend Primer	Term Mix
54144	CTGAAAGACACCATTTAT	CGT
183535	GTCCTGCACAAGACTTTGATA	ACG
236104	CCCATTTCATACCACCTATCA	ACG
236105	CTCCCTCCTCCTTGAGACC	ACT

dbSNP rs#	Extend Primer	Term Mix
236113	GATCATTCATGAAACAGATTCTA	ACG
236119	TGTTCTCAAGGAAAAAAGAAAAAA	ACT
236120	GATCATCCCTGGGAATGGTA	ACT
236122	GAGGCAGGGAATCAGCAATA	ACT
236143	ACCCCAAATAGGTAAAGATCTGT	ACT
236145	CTCCTGCACTGAGCTCCTAT	ACT
236146	TATCTCCGCCCTAAGAATACT	ACT
236149	GAAATATTAGAATTTAGAGGCAG	ACT
236160	GAGTTTTTATGAGAAAGGGCAA	ACT
236162	GTTGTTTTAAAGTGTTGGTTGTAA	ACT
236163	CAATACATAGTGAAGCTTTGGG	ACT
236164	CTGGTGTCACACACACATGTA	ACT
236166	GGAACATCTCAGAAAAAAAA	CGT
236167	CTTGTGCAGGACTTAGACCA	ACT
236176	ATAGGCTTTCTTGTGTATTTGCA	ACT
236177	AGTGAGGGGAAGCAGAGTC	CGT
236180	ACTTGGGAATTAGGTGGAGG	ACG
236182	GTTCAGAGATAATGCTGCTGATC	CGT
236183	GAAATCCCACAGGAACACAAT	ACT
236184	CCCAGCAAATAACAAGAATTGGCC	ACT
379418	CTTAAGCCAAGACAAACA	ACG
400735	TTCATCTTCCACCCTGGCC	ACT
409035	TGCTTTGCTTGCCTCCCACA	ACG
440005	GCTGTCCTTTTTACAAGGAAAT	ACT
446614	TAAAGACTGAAGCTTTCACAGT	ACT
451417	CGTTATCACCATTGGGCTTTA	CGT
454422	GATCCTTCTCACTTACTGTTC	ACT
500277	GATTATGCCCTGAGGTCTTTTG	ACT
540717	AACCCAACAAATCCTAGGGC	ACG
546106	GATGGGCTCCCCATATGAC	ACT
571039	TGTAGCAGGCAACTGAGCAGGAGA	CGT
1039542	GAACACCCTCCAGCACAAG	ACT
1039543	AGAAGAGCTTTCATCTGTGTG	ACT
1343180	CACAGCCCTCCATTACAGC	ACT
CHGB_SNP2	GTATGTTTGTGTTCCATTTGCA	ACT
rs384578	GTATGTTTGTGTTCCATTTGCA	ACT
rs742710	GAAGAGAGGGGCCTTGAGC	ACG
rs742711	TCCCCTGCCTCTGTGATGG	ACG
rs881118	CTGAAGCCTGCTTTCTTTCAT	ACT
rs910122	CTTCACTGCTATCTTCCCCT	ACG
CHGB_SNP1	TCTTGGAGCCCTGTATTC	ACG
180477	GTGGCTCACGCCTATAA	CGT
236102	GTTCTCTCTGAAACCTGTTA	ACT
236103	CATGCACCAGCTGTGTG	ACT

dbSNP rs#	Extend Primer	Term Mix
236106	GAACATTCCAGGCAAAC	ACT
236107	GTTCTGGTAAAAAAAAAGTTTG	CGT
236108	CTGTTGAACCCAGAAATATC	ACT
236109	CACTAGCCACCACGCCC	ACT
236110	CAATGTTGACTGCATTGACT	ACT
236111	GTTCTGAGGTTACCAGA	ACT
236112	ACCATCCTAGCTAACACG	ACT
236114	AATCACAAGTACCTCGAATAC	ACT
236115	AGGTAAGTGGCAGAACT	ACT
236116	TCAGGCAAGCACAGTACAAA	ACG
236117	GCCTCCCAAGTAGCTGG	ACT
236118	CCCTAATCTGAAAATCTGAAATCT	ACT
236121	AAGAATTTTCTTATTCAACTGTC	ACT
236123	CTTCACTAAATAAAAATGTGTCC	ACG
236124	GTTTGGCTCACGGAATTAT	ACG
236125	TTTAACTCCTAGCTTTTAAAGA	ACG
236148	AATGTGGCTGGTCCGATCTG	ACT
236151	ATTCTCCTGGCTCCCTG	ACG
236152	TTATCCAGGCGTCCAGG	ACT
236154	TGATGCCACTGGTCAGG	ACT
236155	AATTCCCCTTTGCACTCAT	ACT
236158	TTACAACTGTAAGCCACCGC	ACT
236159	CCTGACCTCGTGATCTG	ACG
236161	CTCTAGGCTAGTATTAATTTTTGT	ACT
236165	TAACGCCTGTAATCCCA	ACT
236168	ACTTGCATGTGTATGTATATATCT	ACT
236169	ATGTCTTTTCCCCCTCT	ACT
236171	AAGTGCTGGGATTACAGATA	ACG
236173	TTGCTCCCTCTCCCCTT	ACT
236175	CACTGCATCAAGATGGGCC	ACG
236179	CAGGTTTGACTCAAAACTTTAA	ACG
236181	ATGTCTTTTCTGTACTGGATA	ACT
236185	TACTGAGGAGGCTGAGG	ACG
236187	AACTCTGTCTCTGCAAAAAAA	ACT
236189	CAGAAAAACTGCACAAAAA	ACT
364652	CTGTGATAGGAAAAAAGGAA	ACT
394604	CCAGCAGAGGCAAAAATAAGA	ACG
403727	TGCCACTGCACTCCAGCCT	ACT
446658	AGGAAAAGAGAGGCAAAC	ACT
451571	GCTGTCTTCATTCTCTTGT	ACG
452749	CCTATTTTCAAGTCAGGT	ACG
454328	CCTAAACAGCAGTTTTAGTACAT	ACG
742710	AGAGAGGGGCCTTGAGC	ACG
742711	TGAGCCGGGAAAGGGAC	ACT

dbSNP rs#	Extend Primer	Term Mix
881118	TGAAGCCTGCTTTCTTTCAT	ACT
910122	CTTCACTGCTATCTTCCCCT	ACG
1005517	GGTAGAGAATGTAATAACAGT	ACT
1394095	ACGAGAGGGGCGGGGCG	ACT
2206817	TTAGAGCAGGGCAGGGG	ACT
2268339	CAGAATGCTGAGATGGC	ACT
2300427	CACCCGGCCGGGAAAAT	ACT
2326680	TGGAATTTGAGAAGGCCTG	ACT
2423131	GCCTTACGAGAATGTTATTT	CGT
2876003	AGAGTCTGTAGTCCCAA	ACT
3761873	TGTATTTTCCATAGTAATTTGCTC	ACT

Genetic Analysis of Allelotyping Results

[0269] Allelotyping results are shown for cases and controls in Table 21. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where "AF" is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, SNP rs2300427 has the following case and control allele frequencies: case A1 (T) = 0.615; case A2 (G) = 0.385; control A1 (T) = 0.605; and control A2 (G) = 0.395, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped ("not AT").

TABLE 21

dbSNP rs#	Position in Fig 3	Chrom Position	Alleles (A1/A2)	A2 Case AF	A2 Control AF	p-Value
2300427	186	5842586	T/G	0.395	0.385	0.7201
571039	1332	5843732	G/T	0.669	0.695	0.3506
236143	1893	5844293	T/C	0.499	0.545	0.1289
236145	2786	5845186	C/G	0.198	0.188	0.6666
236146	2962	5845362	A/G	0.549	0.601	0.0802
446658	3377	5845777	A/C	0.013	0.011	0.8118
500277	5522	5847922	A/G	0.217	0.214	0.8990
2268339	5621	5848021	A/G	0.931	0.888	0.0128
454328	5889	5848289	G/A	0.330	0.323	0.7912
236148	7531	5849931	T/C	0.463	0.459	0.9081
236149	8268	5850668	T/C	0.550	0.553	0.9289
910122	8923	5851323	G/A	0.373	0.337	0.2228
881118	8988	5851388	A/C	0.026	0.042	0.1278
236151	9117	5851517	G/A	0.295	0.272	0.3986
236152	9448	5851848	C/G	0.310	0.316	0.8329
CHGB_SNP1	9494	5851894	G/A	0.497	0.485	0.6989
742710	9628	5852028	G/A	0.010	0.044	0.0007
742711	9640	5852040	T/C	0.326	0.266	0.0282

dbSNP rs#	Position in Fig 3	Chrom Position	Alleles (A1/A2)	A2 Case AF	A2 Control AF	p-Value
236154	11072	5853472	T/G	0.410	0.421	0.6955
236155	11150	5853550	A/G	0.454	0.445	0.7568
54144	11379	5853779	C/A	0.326	0.350	0.3946
540717	11692	5854092	G/A	0.492	0.495	0.9036
236158	12056	5854456	T/C	0.251	0.229	0.3815
236159	12104	5854504	G/A	0.404	0.385	0.5205
446614	14160	5856560	T/C	0.968	0.951	0.1381
1343180	14836	5857236	A/G	0.376	0.383	0.7937
1039542	14980	5857380	A/C	0.348	0.390	0.1491
400735	15165	5857565	A/G	0.286	0.272	0.5984
1039543	15315	5857715	A/G	0.356	0.318	0.1914
440005	15624	5858024	A/C	0.563	0.575	0.6852
394604	15796	5858196	C/T	0.297	0.277	0.4715
546106	15939	5858339	T/C	0.733	0.714	0.4816
452749	16581	5858981	C/T	0.297	0.262	0.1967
364652	17045	5859445	T/C	0.513	0.558	0.1426
403727	18501	5860901	A/G	0.743	0.748	0.8547
236160	21800	5864200	A/G	0.893	0.794	0.0000
236161	21966	5864366	A/G	0.121	0.187	0.0026
236162	22134	5864534	A/G	0.735	0.648	0.0021
236163	22181	5864581	A/G	0.042	0.096	0.0006
236164	23028	5865428	A/G	0.815	0.734	0.0018
236165	23312	5865712	A/G	0.888	0.892	0.8319
236166	23573	5865973	T/A	0.060	0.125	0.0004
236167	23858	5866258	A/G	0.695	0.675	0.4700
183535	23888	5866288	G/A	0.088	0.149	0.0020
236168	23990	5866390	T/C	0.302	0.303	0.9685
236169	24073	5866473	A/G	0.048	0.115	0.0001
236171	25330	5867730	G/A	0.749	0.739	0.6961
236173	26473	5868873	T/C	0.051	0.119	0.0002
236175	27958	5870358	C/T	0.172	0.289	0.0000
236176	28421	5870821	A/G	0.056	0.143	0.0000
451571	28804	5871204	C/T	0.954	0.924	0.0378
236177	29322	5871722	C/A	0.072	0.125	0.0040
1005517	30819	5873219	G/C	0.970	0.934	0.0051
236179	31956	5874356	G/A	0.056	0.108	0.0021
236180	32592	5874992	G/A	0.073	0.124	0.0047
236181	32818	5875218	G/C	0.908	0.807	0.0000
236182	32880	5875280	G/T	0.914	0.844	0.0006
236183	33244	5875644	G/C	0.870	0.771	0.0001
236184	33845	5876245	A/G	0.059	0.120	0.0006
236185	34272	5876672	G/A	0.028	0.041	0.2547
236187	34931	5877331	T/C	0.048	0.088	0.0090
1394095	36870	5879270	T/G	0.880	0.884	0.8136
236189	37790	5880190	T/C	0.105	0.160	0.0077
236110	38708	5881108	T/G	0.875	0.799	0.0009
236111	39135	5881535	T/C	0.045	0.080	0.0161
236112	39919	5882319	A/G	0.852	0.854	0.9265
236113	40166	5882566	C/T	0.869	0.790	0.0007
236114	40985	5883385	A/G	0.345	0.347	0.9357
236115	41049	5883449	A/G	0.835	0.749	0.0007
236116	41935	5884335	C/T	0.679	0.677	0.9439
236117	42775	5885175	A/C	0.926	0.903	0.1611
236118	43807	5886207	T/C	0.771	0.697	0.0060
236119	44254	5886654	A/G	0.773	0.694	0.0032
3761873	44814	5887214	A/C	0.044	0.042	0.8940
236120	45249	5887649	T/G	0.072	0.109	0.0291
451417	47599	5889999	C/A	0.086	0.103	0.3196

dbSNP rs#	Position in Fig 3	Chrom Position	Alleles (A1/A2)	A2 Case AF	A2 Control AF	p-Value
379418	47807	5890207	G/A	0.102	0.166	0.0022
2326680	48555	5890955	A/C	0.610	0.542	0.0233
409035	49249	5891649	G/A	0.788	0.700	0.0012
454422	49293	5891693	A/C	0.763	0.673	0.0013
236102	57566	5899966	A/C	0.856	0.780	0.0014
236103	63587	5905987	T/C	0.798	0.740	0.0205
236104	64560	5906960	C/T	0.793	0.714	0.0029
236105	65432	5907832	C/G	0.077	0.127	0.0065
236106	66291	5908691	T/C	0.109	0.154	0.0265
236107	71331	5913731	A/T	0.063	0.103	0.0159
180477	73344	5915744	A/T	0.768	0.698	0.0086
236108	74159	5916559	C/T	0.825	0.742	0.0012
236109	74564	5916964	T/C	0.162	0.170	0.7214
236121	78194	5920594	A/G	0.136	0.182	0.0345
236122	79128	5921528	T/C	0.862	0.790	0.0019
236123	79393	5921793	C/T	0.742	0.725	0.5078
236124	81579	5923979	G/A	0.867	0.800	0.0029
236125	82574	5924974	C/T	0.069	0.104	0.0385
2876003	85309	5927709	G/C	0.847	0.836	0.6067
CHGB_SNP2	87076	5929476	A/G	0.783	0.734	0.0538
2423131	87844	5930244	C/A	0.471	0.427	0.1473
2206817	90241	5932641	T/C	0.974	0.985	0.2027

[0270] Figure 16 shows the proximal SNPs in and around the CHGB and C20orf154 gene region. As indicated, some of the SNPs were untyped. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 16 can be determined by consulting Table 18. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0271] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to

the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than 10^{-8} were truncated at that value.

[0272] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link (http address: www.ncbi.nlm.nih.gov/LocusLink/) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is place at the 3' end of each gene to show the direction of transcription.

Additional Genotyping

[0273] In addition to the CHGB/C20orf154 incident SNP, two other SNPs (rs742710 and rs236110) were genotyped in the discovery cohort and found to be significantly associated with breast cancer. See Table 24.

[0274] The methods used to verify and genotype the proximal SNP of Table 21 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 22 and Table 23, respectively.

TABLE 22

dbSNP rs#	First PCR primer	Second PCR primer
742710	ACGTTGGATGGACATGAAATAGGCACGTGG	ACGTTGGATGAGAAAGTGAGGAAGAGAGGG
236110	ACGTTGGATGTCACTCTGTTTCTACTAACC	ACGTTGGATGATCGATCCAATGTTGACTGC

TABLE 23

dbSNP rs#	Extend Primer	Term Mix
742710	AGAGAGGGGCCTTGAGC	ACG
236110	AATGTTGACTGCATTGACT	ACT

[0275] Table 24, below, shows the case and control allele frequencies along with the p-values for the SNPs genotyped. The disease associated allele of column 4 is in bold and the disease associated amino acid of column 5 is also in bold. For rs742710 the proline is associated with breast cancer, and for rs236110, the glutamine is associated with breast cancer. The chromosome position provided corresponds to NCBI's Build 33.

TABLE 24: Genotpying Results

dbSNP rs#	Position in Figure 4	Chromo-some Position	Alleles (A1/A2)	Amino Acid Change	AF F case	AF F control	p-value	Odds Ratio
742710	9628	5852028	G/A	P413L	A = 0.030 G = 0.970	A = 0.060 G = 0.940	**0.0279**	0.51
236110	38708	5881108	T/G	Q63K	T = 0.080 G = 0.920	T = 0.130 G = 0.870	**0.0112**	1.66

Example 6
LOC338749 Proximal SNPs

[0276] It has been discovered that a polymorphic variation (rs763471) in the LOC338749 gene region is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs763471) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. Approximately sixty-three allelic variants located within the LOC338749 region were identified and allelotyped. The polymorphic variants are set forth in Table 25. The chromosome position provided in column four of Table 25 is based on Genome “Build 33” of NCBI’s GenBank.

TABLE 25

dbSNP rs#	Chromosome	Position in Figure 4	Chromosome Position	Allele Variants
2957666	11	142	10442142	A/G
2198010	11	693	10442693	A/C
2198009	11	731	10442731	T/C
2198008	11	879	10442879	T/G
2957667	11	1084	10443084	C/A
2957669	11	2249	10444249	C/G
2957670	11	2519	10444519	A/G
2923115	11	4461	10446461	G/A
2957679	11	4616	10446616	A/G
2957678	11	5109	10447109	G/A
2957677	11	5270	10447270	C/G
2923117	11	5436	10447436	G/C
2957675	11	5457	10447457	T/C
750371	11	6536	10448536	G/C
1562781	11	9665	10451665	T/A
752373	11	16120	10458120	C/T
3741045	11	29489	10471489	T/C
3741044	11	29524	10471524	G/T
763470	11	49159	10491159	A/G

dbSNP rs#	Chromosome	Position in Figure 4	Chromosome Position	Allele Variants
763471	11	49273	10491273	G/T
1376001	11	49596	10491596	A/G
1376000	11	50135	10492135	A/G
1450274	11	50184	10492184	C/G
1375999	11	50393	10492393	A/C
1450273	11	50401	10492401	T/G
1450270	11	55750	10497750	T/C
899013	11	73843	10515843	T/C
2071019	11	73852	10515852	T/C
LOC_SNP1	11	74052	10516052	G/A
LOC_SNP2	11	75382	10517382	A/G
930672	11	75662	10517662	T/A
922359	11	75942	10517942	T/C
2403330	11	77917	10519917	A/G
936513	11	78821	10520821	A/G
3891547	11	94813	10536813	C/G
3741043	11	97149	10539149	T/C

Assay for Verifying and Allelotyping SNPs

[0277] The methods used to verify and allelotype the proximal SNPs of Table 25 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 26 and Table 27, respectively.

TABLE 26

dbSNP rs#	Forward PCR primer	Reverse PCR primer
752373	ACGTTGGATGTATCACAAGAACAGCATGGG	ACGTTGGATGATGGTTTCTGTAATCCCCCC
763470	ACGTTGGATGAAGAGGAGTGGCTGATAATG	ACGTTGGATGAAGCAGAAAACTTTGTGCCG
763471	ACGTTGGATGGGCAGGTCATGGATTTATTG	ACGTTGGATGCATCATTCCTCTGTGAGGCG
763471	ACGTTGGATGGTGAAGAGCTCTGAAATGCC	ACGTTGGATGTAACTCCTGTGTGGCTTTCT
899011	ACGTTGGATGAAGGTGGAGCCTGCCTCAAG	ACGTTGGATGAGCTTTGCACCCTGTGATGC
899011	ACGTTGGATGAAGGTGGAGCCTGCCTCAAG	ACGTTGGATGAGCTTTGCACCCTGTGATGC
922359	ACGTTGGATGTCAAGCGATCCTCTTCAGCC	ACGTTGGATGATTCATTCCAAGACCGGGTG
930672	ACGTTGGATGGTGGGTTACTTGGTCCATAC	ACGTTGGATGACAGAGCAAGACCTTCTCTC
936513	ACGTTGGATGGTATGAAGTTCTTTGCAGAGT	ACGTTGGATGTACTACTGCACTCCAGCCTG
1375999	ACGTTGGATGCCATTCTTTTACCTTGAACC	ACGTTGGATGCAGAGACTTGCAGAATGGAC
1376000	ACGTTGGATGATAGCTGATGGTGTGCTGAG	ACGTTGGATGAAGCTTGCCTCCCAAGTTAG
1376001	ACGTTGGATGCAAACAATCCCATTACACAG	ACGTTGGATGCAGTACAACAGGGTGGCTATC
1450270	ACGTTGGATGCCATATCACATGGATATGAGG	ACGTTGGATGCATGGCTTCTCTTACACCTG
1450273	ACGTTGGATGGCTGCATATAAGAGACACATG	ACGTTGGATGGCCACTCCAGCTTTCTTTTG
1450274	ACGTTGGATGTGAGAGGAAGCCTGGTGTTG	ACGTTGGATGAAGCTTGCCTCCCAAGTTAG

dbSNP rs#	Forward PCR primer	Reverse PCR primer
1562781	ACGTTGGATGTATGTCTCCTGCCTTCTTCC	ACGTTGGATGGAAAGAAGCTTGATGTGGC
2071019	ACGTTGGATGGGTAAACAACTGACCCATCC	ACGTTGGATGCCTGGGAAATAACCATGAGC
2071020	ACGTTGGATGAATTCACAGCTAAGCCTCCC	ACGTTGGATGTTCAGCTCCAGCTGCATGTT
2198008	ACGTTGGATGGTAGAAGTTTAGTATATGATG	ACGTTGGATGCCCTGTCATTTCAAATACCG
2198009	ACGTTGGATGCTTGTGCCAATCCCACAATG	ACGTTGGATGGCAGAAGTCTAGCCAAGAAC
2198010	ACGTTGGATGCTTGTGCCAATCCCACAATG	ACGTTGGATGAATGCAGAAGTCTAGCCAAG
2403330	ACGTTGGATGTAACTCTGAGACCCAAGGAC	ACGTTGGATGCCAGACAGTTGTGTGTTGAC
2923115	ACGTTGGATGGGATTACCCTAAGGATCCAC	ACGTTGGATGAGAGGAATTCAGTTGCTGCC
2923117	ACGTTGGATGTTGAGTCCAAGAGGTTGAGG	ACGTTGGATGAGACAGTCTTGCTCTGTCAC
2957666	ACGTTGGATGTACTTGGGAGACTGAGGTAG	ACGTTGGATGGCATAGTGGTGTGATCATGG
2957667	ACGTTGGATGAGAATGGTCTTTCCCACTCC	ACGTTGGATGATGGATTACGGAAGGAATAC
2957669	ACGTTGGATGTACTGAGACTCCCAGCATTG	ACGTTGGATGGTGTGCAGCTTAGTAAGTGC
2957670	ACGTTGGATGTCATGTGATTCTCCTGCCTC	ACGTTGGATGGTGAAACCCCGTCTCTACTA
2957675	ACGTTGGATGAGAATGACTTGGGTTTTGGG	ACGTTGGATGCAGTGAGTTGTGACAGCACC
2957677	ACGTTGGATGGTCTTTCTCAATCCCAGCAC	ACGTTGGATGACGAGATCTCCTTGTGTTGC
2957678	ACGTTGGATGAAGACCTCAGGATGTGATGC	ACGTTGGATGATGACCCCGTTTCTTTGCAC
2957679	ACGTTGGATGAGTTCGTCAGAGAGATGTCC	ACGTTGGATGGAGCACATGGATTCACAGAG
3741043	ACGTTGGATGGACATCAGAAGCTAATTGGG	ACGTTGGATGCTTCTTAATGGTAGGGCCAG
3741044	ACGTTGGATGTTTGTTATGCAGAGGTGGCC	ACGTTGGATGTAGATGGGCTCTTCTTGGAC
3741045	ACGTTGGATGAACTGAGCTTCAGACTTCCC	ACGTTGGATGTCAGACCTGTAGATGGGCTC
3891547	ACGTTGGATGGCCATCAAGTTTGTGGCAAT	ACGTTGGATGAAGCTATATGGAGCCCAAGG

TABLE 27

dbSNP rs#	Extend Primer	Term Mix
752373	GGGACCAGGTGGAGATAA	ACG
763470	AGAAAACTTTGTGCCGTTTTCT	ACT
763471	CCAGGCAGCAACTCCCT	ACT
763471	CTCCAAGCAGTAAAGATGTTC	CGT
899011	TTGGTTTTAGAGGATTGCTCC	ACG
899011	GGTTTTAGAGGATTGCTCC	ACG
922359	TCATGCCTATAATCCAAGCA	ACT
930672	AAAAGCAAGAAACAACAGCA	CGT
936513	AGACAGGGTGAGACCTC	ACT
1375999	TATGCTGCATATAAGAGACACAT	ACT
1376000	ACATATTTCTGGTCTCCA	ACT
1376001	ACAGGGTGGCTATCATTAAC	ACT
1450270	TGTGAACTGAAAAGTCAAG	ACT
1450273	CTTGAACCTATTTCTGTTTTT	ACT
1450274	CTCCCAAGTTAGATTGGTTA	ACT
1562781	CTTGATGTGGCTGAAGT	CGT
2071019	GTGCTGTTGAAATCCTGGG	ACT

2071020	AACCCTTTGTCAGCTGAA	ACT
2198008	CAGGCTTTTGGCTAAGATCAAG	ACT
2198009	AGTGAAGAATTTTCCCTATTAGAT	ACT
2198010	AAGTCTAGCCAAGAACATTT	ACT
2403330	CTCCCACTCCTCTCATCAG	ACT
2923115	GTTGCTGCCCGCTTTCC	ACG
2923117	CTCTGTCACCCATGCTGGA	ACT
2957666	ACCTCCTGGGCTCAAGC	ACT
2957667	GGAAGGAATACTAAAGAACAA	CGT
2957669	AGTAAGTGCTGTGATGCACC	ACT
2957670	TAGCTGAGCATGGTGGC	ACT
2957675	AGCATGGGTGACAGAGC	ACT
2957677	TGTGTTGCCCAGACTAG	ACT
2957678	ACTCCCTGGCCTCCCCT	ACG
2957679	TCACAGAGCTGCCAGGG	ACT
3741043	CAAAATTCTCCTGCCAC	ACT
3741044	GGGGAAAGGGAAGTCTG	CGT
3741045	TAGATGGGCTCTTCTTG	ACT
3891547	TATGGAGCCCAAGGATGACC	ACT

Genetic Analysis of Allelotyping Results

[0278] Allelotyping results are shown for cases and controls in Table 28. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where "AF" is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP rs2957666 has the following case and control allele frequencies: case A1 (A) = 0.863; case A2 (G) = 0.137; control A1 (A) = 0.855; and control A2 (G) = 0.145, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped.

TABLE 28

dbSNP rs#	Position in Figure 4	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
2957666	142	10442142	A/G	0.145	0.137	0.6913
2198010	693	10442693	A/C	0.424	0.410	0.6335
2198009	731	10442731	T/C	0.485	0.469	0.6093
2198008	879	10442879	T/G	0.424	0.422	0.9461
2957667	1084	10443084	C/A	0.437	0.430	0.8105
2957669	2249	10444249	C/G	0.556	0.552	0.8955
2957670	2519	10444519	A/G	0.051	0.054	0.8305
2923115	4461	10446461	G/A	0.490	0.475	0.6091
2957679	4616	10446616	A/G	0.391	0.403	0.6947
2957678	5109	10447109	G/A	0.458	0.447	0.7210
2957677	5270	10447270	C/G	0.479	0.475	0.9147

dbSNP rs#	Position in Figure 4	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
2923117	5436	10447436	G/C	0.913	0.950	0.0164
2957675	5457	10447457	T/C	0.425	0.421	0.8857
750371	6536	10448536	G/C	0.143	0.174	0.1547
1562781	9665	10451665	T/A	0.966	0.973	0.4889
752373	16120	10458120	C/T	0.251	0.195	0.0255
3741045	29489	10471489	T/C	0.715	0.719	0.8787
3741044	29524	10471524	G/T	0.419	0.374	0.1272
763470	49159	10491159	A/G	0.139	0.137	0.9405
763471	49273	10491273	G/T	0.481	0.537	0.0661
1376001	49596	10491596	A/G	0.538	0.484	0.0705
1376000	50135	10492135	A/G	0.228	0.256	0.2849
1450274	50184	10492184	C/G	0.777	0.736	0.1185
1375999	50393	10492393	A/C	0.933	0.913	0.2078
1450273	50401	10492401	T/G	0.845	0.817	0.2102
1450270	55750	10497750	T/C	0.349	0.264	0.0025
899013	73843	10515843	T/C	0.388	0.385	0.9145
2071019	73852	10515852	T/C	0.824	0.782	0.0756
LOC_SNP1	74052	10516052	G/A	0.245	0.217	0.2760
LOC_SNP2	75382	10517382	A/G	0.283	0.327	0.1112
930672	75662	10517662	T/A	0.157	0.154	0.8763
922359	75942	10517942	T/C	0.106	0.120	0.4749
2403330	77917	10519917	A/G	0.624	0.574	0.0959
936513	78821	10520821	A/G	0.381	0.423	0.1554
3891547	94813	10536813	C/G	0.069	0.079	0.5157
3741043	97149	10539149	T/C	0.855	0.832	0.2972

[0279] Figure 17 shows the proximal SNPs in and around the *LOC338749* region for females. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 17 can be determined by consulting Table 28. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0280] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit

test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than 10^{-8} were truncated at that value.

[0281] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link (http address: www.ncbi.nlm.nih.gov/LocusLink/) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is place at the 3' end of each gene to show the direction of transcription.

Example 7

TTN/LOC351327 Proximal SNPs

[0282] It has been discovered that a polymorphic variation (rs2046778) in the TTN/LOC351327 region is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs2046778) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. Approximately forty-six allelic variants located within the TTN/LOC351327 region were identified and allelotyped. The polymorphic variants are set forth in Table 29. The chromosome position provided in column four of Table 29 is based on Genome "Build 33" of NCBI's GenBank.

TABLE 29

dbSNP rs#	Chromosome	Position in Figure 5	Chromosome Position	Allele Variants
2291309	2	200	179587600	T/G
2291310	2	381	179587781	T/C
1484119	2	5303	179592703	G/C
TTN_SNP1	2	6084	179593484	C/T
1484120	2	6879	179594279	T/A
2291312	2	7837	179595237	T/C
3816782	2	7985	179595385	C/A
2291313	2	9333	179596733	T/C
2306636	2	11559	179598959	T/C
2291304	2	12473	179599873	T/C
2291305	2	12880	179600280	T/A
1905520	2	13606	179601006	C/T
2291306	2	14861	179602261	A/G
TTN_SNP2	2	20658	179608058	C/T
2054708	2	22200	179609600	G/A
2306637	2	24525	179611925	A/C
3769863	2	26373	179613773	T/G

dbSNP rs#	Chromosome	Position in Figure 5	Chromosome Position	Allele Variants
3769860	2	42869	179630269	A/T
3816849	2	43713	179631113	A/G
3769858	2	44429	179631829	A/G
2279472	2	49037	179636437	A/G
2046778	2	49170	179636570	A/G
1565288	2	50206	179637606	G/A
2129108	2	51552	179638952	C/T
2170850	2	51674	179639074	T/G
2029397	2	56427	179643827	T/C
2029395	2	56844	179644244	G/A
1844334	2	57953	179645353	A/G
998329	2	60862	179648262	G/A
1489486	2	61606	179649006	T/C
2046777	2	62560	179649960	G/A
1489483	2	65078	179652478	A/G
1489482	2	65155	179652555	G/T
2366911	2	70295	179657695	T/C
2366912	2	70335	179657735	G/T
2366913	2	70398	179657798	C/T
2078403	2	79233	179666633	C/T
1489481	2	80025	179667425	C/G
2129111	2	84521	179671921	A/G
966783	2	84540	179671940	C/T
1489480	2	85170	179672570	T/G
1489479	2	85300	179672700	A/C
726215	2	87596	179674996	A/C
1387472	2	89696	179677096	C/A
2086832	2	92219	179679619	A/T
1872203	2	96589	179683989	A/T

Assay for Verifying and Allelotyping SNPs

[0283] The methods used to verify and allelotype the proximal SNPs of Table 29 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 30 and Table 31, respectively. The methods used to verify and allelotype the proximal SNPs of Table 29 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 30 and Table 31, respectively.

TABLE 30

dbSNP rs#	Forward PCR primer	Reverse PCR primer
726215	ACGTTGGATGAACTGAGCCCCATGAAATGC	ACGTTGGATGAAAACAGCAATTGAGAACAC
966783	ACGTTGGATGCTCCTGAATTTTAGCCATAC	ACGTTGGATGTACGCAATAGTTCCTGGGAG
998329	ACGTTGGATGGAAGAGCACATTATTTGCTGG	ACGTTGGATGACACACTGGTGTTTTGTCAG
1387472	ACGTTGGATGGAAAGGCCTTGAATTGGAAC	ACGTTGGATGGTTCTGCTAGTGTCATCTTC
1484119	ACGTTGGATGGCAGCTACAATCATAAAGGG	ACGTTGGATGTGTGCCCTTAATAATGGTTG
1484120	ACGTTGGATGATGGTCATGGCATCCAGTTC	ACGTTGGATGGGCTGGTTTCTGACACTATC
1489479	ACGTTGGATGGGAGCATCAGTCATTTTGGG	ACGTTGGATGCACCAGGACATAACATGACG
1489480	ACGTTGGATGGGGTTGTGGAGAATCATTAC	ACGTTGGATGGGTGGCAGTAATCTTCACTT
1489481	ACGTTGGATGTCTCTGCAGTTGAGGAGATG	ACGTTGGATGTTGGGAAAGGCCATCAAGTC
1489482	ACGTTGGATGCTCTGGATAAAAGACTCAGC	ACGTTGGATGCCCTTCCAACAGCTATCTGG
1489483	ACGTTGGATGTTGGTTTGCTATCAATGAAG	ACGTTGGATGGATAGGTGTACACATATAGC
1489486	ACGTTGGATGAAAAAACACACCACAGCCCC	ACGTTGGATGCTTCGTATTTGGCTCTGACC
1552280	ACGTTGGATGATGAAAAGTGACACCCATCC	ACGTTGGATGTCTGAAGCTGTTGAATCAGG
1565288	ACGTTGGATGTAGCCAATTGGTGAACACTC	ACGTTGGATGCTGCCAGTCATAAGGCAAAG
1844334	ACGTTGGATGGCCAAGGAAACTAATTCCTG	ACGTTGGATGCACTTTGGAAGACAGTTCGG
1872203	ACGTTGGATGGTTGCATTAGCTGTTATTCTC	ACGTTGGATGCCAGCAATTCTATTTCAGAG
1905520	ACGTTGGATGCATGGTTTATACTTACTTACG	ACGTTGGATGGTTTATTCCTGTTTCCACAC
2029395	ACGTTGGATGGGAGGGAGACAAAGATTCAC	ACGTTGGATGGCAACAGTTTCACCTTTGGC
2029397	ACGTTGGATGCTCACAGTCCTGAAGACTTG	ACGTTGGATGTGGAAGTGAAGGAGAGAAGC
2046777	ACGTTGGATGGGACTTCAAATATGGTTCAC	ACGTTGGATGTTAAGCCTGGGACTTTTGGG
2046778	ACGTTGGATGGTTCCCTTCCCCCATAAAAC	ACGTTGGATGCATGAAGCCTTATGCTTGAG
2054708	ACGTTGGATGCTAGGCATATCATGCCTCTG	ACGTTGGATGTTGAGCTCACTGTTACCTGC
2078403	ACGTTGGATGTGTGCTCAGGATCGACAGAC	ACGTTGGATGACTCGAGACAACCTACAAGG
2086832	ACGTTGGATGCTTTTGAGCATCACATTCCTC	ACGTTGGATGTGCCTAAGCACTGTATAACC
2129108	ACGTTGGATGAACTCCCAGTAAGTCCTTCC	ACGTTGGATGACTCAGGCAGTAACTCCAAC
2129111	ACGTTGGATGTACACTTTTCCCGCAAGACC	ACGTTGGATGGTCATGGACATCTACAGTATC
2170850	ACGTTGGATGGAAGGCCAATGCAAGGATAC	ACGTTGGATGAAGAACACACAAAAAAAT
2279472	ACGTTGGATGGAGAAGAGCATTGGTTGCTG	ACGTTGGATGTGCCCACAAGTGCTATCTAC
2291304	ACGTTGGATGGTCTCAGGAAGGTTTAGAGG	ACGTTGGATGAAAAGACAAACGATATGGCC
2291305	ACGTTGGATGCATGATTTCAAAATCATGTTC	ACGTTGGATGGAGATGTACAGTATGAGTCC
2291306	ACGTTGGATGCAGCGACTAGTCATTAACCG	ACGTTGGATGCAGTTGGTTTCAACTCTGCC
2291309	ACGTTGGATGCATTGTTGTTCCTACCATTC	ACGTTGGATGAAAGTGGTAAAGGAGAGGCG
2291310	ACGTTGGATGGTGCTTGATACTTGGCCTAC	ACGTTGGATGCAACTGGAAATTGCCGAAGC
2291311	ACGTTGGATGTCAACATTTACTCCTAGCTC	ACGTTGGATGATTTTGGGCTGTGGTCTTCC
2291312	ACGTTGGATGTGTATTCTCCTGCATCGCTC	ACGTTGGATGTCCAAGTTCAAGAACGACAC
2291313	ACGTTGGATGTTCGAGTTTACCGTATGGTG	ACGTTGGATGGATCACAGACAGGTCAGTTG
2306636	ACGTTGGATGCTGAGACCAGTCTGTGTTTG	ACGTTGGATGGTTTCCCATGACACTGTTCC
2306637	ACGTTGGATGCTACTACTATTTCTGGAGTC	ACGTTGGATGCTTATGCATTTCAACTGCCAC
2366911	ACGTTGGATGGTAGATGCTTGAATCAATAAAG	ACGTTGGATGATAGCAGCTCCAGAACTAGG
2366912	ACGTTGGATGGAACTGTTGTTGAATGGGAC	ACGTTGGATGCAATACTTGTAAAATAGCAGC
2366913	ACGTTGGATGCTATCTGTATTCTCATGGCTG	ACGTTGGATGTTACCTAGTTCTGGAGCTGC
3769858	ACGTTGGATGCTACATGTCCATGGTTTGATG	ACGTTGGATGGCATCAACCTTTATGCCAAG
3769860	ACGTTGGATGGTATACAGAATATTGCATGCC	ACGTTGGATGGAACATCATTGAAGGTAAAG
3769863	ACGTTGGATGCAAGGATTTATTACATGCTG	ACGTTGGATGGTCATCAGGAGAAAGTAAGC

dbSNP rs#	Forward PCR primer	Reverse PCR primer
3816782	ACGTTGGATGGAGGAAACCAGAGCTTCAAG	ACGTTGGATGCAGCACGCTGTTTCTCAATG
3816849	ACGTTGGATGAACCAGCTCACCTCAGGAAC	ACGTTGGATGTTTGTGGTGCCCATTCAAAC

TABLE 31

dbSNP rs#	Extend Primer	Term Mix
726215	TTGAGAACACAGGATGC	ACT
966783	CTCCCATTTTGGTCTTG	ACG
998329	GGTGTTTTGTCAGTACAATT	ACG
1387472	ACTACAAACTCTTCCTTACC	CGT
1484119	GTTGTTTATGTTATGTTATGTGTT	ACT
1484120	TGTGCCTCAGTTTCTCC	CGT
1489479	GACAGCTGTAATTGTAGACC	ACT
1489480	CTCAATCACATTTACCCTC	ACT
1489481	TCTGATTGTTCCATTAATATCTG	ACT
1489482	CAGCTATCTGGAAATCTTGTTTGA	CGT
1489483	GTGTACACATATAGCAACCTCA	ACT
1489486	CTCTGACCTGTGAGCTAC	ACT
1552280	GCTGTTGAATCAGGATTTGATT	ACG
1565288	GGCAAAGAAACACTAGAAA	ACG
1844334	CAGTTCGGCAGTTTCTT	ACT
1872203	AAAAAATCATGAAAAGGAGCATG	CGT
1905520	ACAAGTCTTTTCATGGTC	ACG
2029395	CAAAATGAAGGAACACTTATCA	ACG
2029397	AGCTCTGTTGGCACTTT	ACT
2046777	GAGCCTGATTATTTGTTTGGGTA	ACG
2046778	CTGTCATGATTGACAGGTCC	ACT
2054708	CCTGGGCCTGGAAGGCAAC	ACG
2078403	GGCTGGAGCAAGAATTA	ACG
2086832	CAATGTAATCCTTGGATAGAT	CGT
2129108	CAACTACATAGTCAGACTTT	ACT
2129111	TATACGCAATAGTTCCTGGG	ACT
2170850	GAACACACAAAAAAATTTAATCA	ACT
2279472	CTCTTTAAACCTGCATTTTC	ACT
2291304	CGATATGGCCATTTTGG	ACT
2291305	CATATTCACACAATGGGAAAA	CGT
2291306	CTGCCAACTATCAGCTT	ACT
2291309	GCGAGACCATGGCATATAACA	ACT
2291310	AACTTACACGTTTGTTGCTA	ACT
2291311	GTGGTCTTCCGGATATCA	ACG
2291312	CGACACAAATATGTAGTGGA	ACT
2291313	TGTCTTGCTACATTCCAGT	ACT
2306636	TCCAGTAAAATGGTTCCATAAGA	ACT

dbSNP rs#	Extend Primer	Term Mix
2306637	TCAACTGCCACAAAATG	ACT
2366911	TTCCTTTGTCCCATTCA	ACT
2366912	TGTAAAATAGCAGCTCCAGAA	CGT
2366913	ATTCTAAATGGAAAAAGAGCCA	ACG
3769858	TGCCCTGAATGTGCCTC	ACT
3769860	GGATAAGCATATGTAACTTTACG	CGT
3769863	AAGTAAAAAGGACATAAAAACCT	ACT
3816782	GTTGATGGAACAACATAAAA	CGT
3816849	GCCCATTCAAACATAAAG	ACT

Genetic Analysis of Allelotyping Results

[0284] Allelotyping results are shown for cases and controls in Table 32. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where "AF" is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP rs2291309 has the following case and control allele frequencies: case A1 (T) = 0.190; case A2 (G) = 0.810; control A1 (T) = 0.215; and control A2 (G) = 0.785, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped.

TABLE 32

dbSNP rs#	Position in Figure 5	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
2291309	200	179587600	T/G	0.785	0.810	0.3114
2291310	381	179587781	T/C	0.069	0.073	0.8256
1484119	5303	179592703	G/C	0.726	0.714	0.6532
TTN_SNP1	6084	179593484	C/T	0.795	0.796	0.9735
1484120	6879	179594279	T/A	0.952	0.957	0.6579
2291312	7837	179595237	T/C	0.064	0.062	0.8817
3816782	7985	179595385	C/A	0.853	0.858	0.8139
2291313	9333	179596733	T/C	0.702	0.662	0.1474
2306636	11559	179598959	T/C	0.921	0.929	0.6434
2291304	12473	179599873	T/C	0.050	0.025	0.0307
2291305	12880	179600280	T/A	0.040	0.037	0.7733
1905520	13606	179601006	C/T	0.014	0.027	0.1559
2291306	14861	179602261	A/G	0.805	0.839	0.1348
TTN_SNP2	20658	179608058	C/T	0.773	0.821	0.0491
2054708	22200	179609600	G/A	0.893	0.883	0.6092
2306637	24525	179611925	A/C	0.999	0.956	0.0000
3769863	26373	179613773	T/G	0.382	0.409	0.3545
3769860	42869	179630269	A/T	0.025	0.022	0.7358
3816849	43713	179631113	A/G	0.374	0.390	0.5938
3769858	44429	179631829	A/G	0.515	0.561	0.1296
2279472	49037	179636437	A/G	0.916	0.871	0.0156
2046778	49170	179636570	A/G	0.171	0.231	0.0122
1565288	50206	179637606	G/A	0.978	0.979	0.9096
2129108	51552	179638952	C/T	0.886	0.807	0.0005

dbSNP rs#	Position in Figure 5	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
2170850	51674	179639074	T/G	0.906	0.863	0.0258
2029397	56427	179643827	T/C	0.094	0.114	0.2585
2029395	56844	179644244	G/A	0.825	0.810	0.5200
1844334	57953	179645353	A/G	0.536	0.533	0.9266
998329	60862	179648262	G/A	0.035	0.055	0.1078
1489486	61606	179649006	T/C	0.646	0.674	0.3300
2046777	62560	179649960	G/A	0.910	0.875	0.0573
1489483	65078	179652478	A/G	0.188	0.205	0.4857
1489482	65155	179652555	G/T	0.058	0.084	0.0944
2366911	70295	179657695	T/C	0.298	0.268	0.2753
2366912	70335	179657735	G/T	0.238	0.186	0.0341
2366913	70398	179657798	C/T	0.259	0.214	0.0772
2078403	79233	179666633	C/T	0.551	0.564	0.6795
1489481	80025	179667425	C/G	0.270	0.251	0.4774
2129111	84521	179671921	A/G	0.324	0.256	0.0127
966783	84540	179671940	C/T	0.210	0.178	0.1714
1489480	85170	179672570	T/G	0.426	0.398	0.3465
1489479	85300	179672700	A/C	0.414	0.372	0.1553
726215	87596	179674996	A/C	0.141	0.194	0.0173
1387472	89696	179677096	C/A	0.106	0.097	0.6242
2086832	92219	179679619	A/T	0.211	0.298	0.0014
1872203	96589	179683989	A/T	0.652	0.577	0.0113

[0285] Figure 18 shows the proximal SNPs in and around the *TTN* region for females. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 18 can be determined by consulting Table 32. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0286] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to

the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than 10^{-8} were truncated at that value.

[0287] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link (http address: www.ncbi.nlm.nih.gov/LocusLink/) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is place at the 3' end of each gene to show the direction of transcription.

Example 8
In Vitro Production of Target Polypeptides

[0288] cDNA is cloned into a pIVEX 2.3-MCS vector (Roche Biochem) using a directional cloning method. A cDNA insert is prepared using PCR with forward and reverse primers having 5' restriction site tags (in frame) and 5-6 additional nucleotides in addition to 3' gene-specific portions, the latter of which is typically about twenty to about twenty-five base pairs in length. A Sal I restriction site is introduced by the forward primer and a Sma I restriction site is introduced by the reverse primer. The ends of PCR products are cut with the corresponding restriction enzymes (*i.e.*, Sal I and Sma I) and the products are gel-purified. The pIVEX 2.3-MCS vector is linearized using the same restriction enzymes, and the fragment with the correct sized fragment is isolated by gel-purification. Purified PCR product is ligated into the linearized pIVEX 2.3-MCS vector and *E. coli* cells transformed for plasmid amplification. The newly constructed expression vector is verified by restriction mapping and used for protein production.

[0289] *E. coli* lysate is reconstituted with 0.25 ml of Reconstitution Buffer, the Reaction Mix is reconstituted with 0.8 ml of Reconstitution Buffer; the Feeding Mix is reconstituted with 10.5 ml of Reconstitution Buffer; and the Energy Mix is reconstituted with 0.6 ml of Reconstitution Buffer. 0.5 ml of the Energy Mix was added to the Feeding Mix to obtain the Feeding Solution. 0.75 ml of Reaction Mix, 50 μl of Energy Mix, and 10 μg of the template DNA is added to the *E. coli* lysate.

[0290] Using the reaction device (Roche Biochem), 1 ml of the Reaction Solution is loaded into the reaction compartment. The reaction device is turned upside-down and 10 ml of the Feeding Solution is loaded into the feeding compartment. All lids are closed and the reaction device is loaded into the RTS500 instrument. The instrument is run at 30°C for 24 hours with a stir bar speed of 150 rpm. The pIVEX 2.3 MCS vector includes a nucleotide sequence that encodes six consecutive histidine amino acids on the C-terminal end of the target polypeptide for the purpose of protein purification. Target

polypeptide is purified by contacting the contents of reaction device with resin modified with Ni^{2+} ions. Target polypeptide is eluted from the resin with a solution containing free Ni^{2+} ions.

Example 9
Cellular Production of Target Polypeptides

[0291] Nucleic acids are cloned into DNA plasmids having phage recombination cites and target polypeptides are expressed therefrom in a variety of host cells. Alpha phage genomic DNA contains short sequences known as attP sites, and *E. coli* genomic DNA contains unique, short sequences known as attB sites. These regions share homology, allowing for integration of phage DNA into *E. coli* via directional, site-specific recombination using the phage protein Int and the *E. coli* protein IHF. Integration produces two new att sites, L and R, which flank the inserted prophage DNA. Phage excision from *E. coli* genomic DNA can also be accomplished using these two proteins with the addition of a second phage protein, Xis. DNA vectors have been produced where the integration/excision process is modified to allow for the directional integration or excision of a target DNA fragment into a backbone vector in a rapid *in vitro* reaction (Gateway™ Technology (Invitrogen, Inc.)).

[0292] A first step is to transfer the nucleic acid insert into a shuttle vector that contains attL sites surrounding the negative selection gene, ccdB (*e.g.* pENTER vector, Invitrogen, Inc.). This transfer process is accomplished by digesting the nucleic acid from a DNA vector used for sequencing, and to ligate it into the multicloning site of the shuttle vector, which will place it between the two attL sites while removing the negative selection gene ccdB. A second method is to amplify the nucleic acid by the polymerase chain reaction (PCR) with primers containing attB sites. The amplified fragment then is integrated into the shuttle vector using Int and IHF. A third method is to utilize a topoisomerase-mediated process, in which the nucleic acid is amplified via PCR using gene-specific primers with the 5' upstream primer containing an additional CACC sequence (*e.g.*, TOPO® expression kit (Invitrogen, Inc.)). In conjunction with Topoisomerase I, the PCR amplified fragment can be cloned into the shuttle vector via the attL sites in the correct orientation.

[0293] Once the nucleic acid is transferred into the shuttle vector, it can be cloned into an expression vector having attR sites. Several vectors containing attR sites for expression of target polypeptide as a native polypeptide, N-fusion polypeptide, and C-fusion polypeptides are commercially available (*e.g.*, pDEST (Invitrogen, Inc.)), and any vector can be converted into an expression vector for receiving a nucleic acid from the shuttle vector by introducing an insert having an attR site flanked by an antibiotic resistant gene for selection using the standard methods described above. Transfer of the nucleic acid from the shuttle vector is accomplished by directional recombination using Int, IHF, and Xis (LR clonase). Then the desired sequence can be transferred to an expression vector by carrying out a

one hour incubation at room temperature with Int, IHF, and Xis, a ten minute incubation at 37°C with proteinase K, transforming bacteria and allowing expression for one hour, and then plating on selective media. Generally, 90% cloning efficiency is achieved by this method. Examples of expression vectors are pDEST 14 bacterial expression vector with att7 promoter, pDEST 15 bacterial expression vector with a T7 promoter and a N-terminal GST tag, pDEST 17 bacterial vector with a T7 promoter and a N-terminal polyhistidine affinity tag, and pDEST 12.2 mammalian expression vector with a CMV promoter and neo resistance gene. These expression vectors or others like them are transformed or transfected into cells for expression of the target polypeptide or polypeptide variants. These expression vectors are often transfected, for example, into murine-transformed a adipocyte cell line 3T3-L1, (ATCC), human embryonic kidney cell line 293, and rat cardiomyocyte cell line H9C2.

Example 10

Haplotype analysis of the *GP6* locus

[0294] rs1671152 is significantly associated with breast cancer at the allele level ($P < 0.05$). This relationship does not hold at the genotype level. Very weak LD is observed across markers in the region. Estimated chi-squared statistics, odds ratios, and score tests indicate that haplotypes are not significantly associated with breast cancer.

Statistics

[0295] Chi-squared statistics are estimated to assess whether 1) alleles and genotypes are associated with breast cancer status and 2) marker genotype frequencies deviate significantly from Hardy-Weinberg equilibrium (HWE). Haplotype frequencies and relative frequencies are estimated, as well as several statistics (r^2, D', and p-value) that gauge the extent and stability of linkage disequilibrium between markers in each region. Chi-squared statistics and score tests are estimated to determine whether reconstructed haplotypes are significantly associated with breast cancer status ($P < 0.05$). P-values are estimated for 1) the full set of reconstructed haplotypes and 2) a reduced set that excludes haplotypes with observed frequencies less than 10. Results are presented by chromosome order.

Results

Summary Statistics: Alleles and Genotypes

SNP Locations

SNP.ID	Type	Location
269911	Proximal	60156885
1671152	Incident	60202366
2124090	Proximal	60240477

Allele by GYNGroup

	N	Case (N=508)	Control (N-536)	Test Statistic
269911 : T	1008	18% (90)	16% (85)	Chi-square=0.77 d.f.=1 P=0.38
1671152 : G	1008	86% (408)	81% (431)	Chi-square=3.98 d.f.=1 P=0.0462
2124090 : A	1022	13% (64)	11% (56)	Chi-square=1.47 d.f.=1 P=0.226

Genotype by GYNGroup

	N	Case (N=254)	Control (N=268)	Test Statistic
269911 : AA	504	68% (165)	71% (184)	Chi-square=0.73 d.f.=2 P=0.693
AT		28% (68)	26% (67)	
TT		5% (11)	3% (9)	
1671152 : TT	504	4% (10)	5% (13)	Chi-square=4.79 d.f.=2 P=0.0912
TG		20% (48)	28% (75)	
GG		76% (180)	67% (178)	
2124090 : CC	511	75% (184)	80% (213)	Chi-square=3.42 d.f.=2 P=0.181
CA		24% (60)	18% (48)	
AA		1% (2)	2% (4)	

Genotype QC: Test of Hardy-Weinberg Proportions

All

	A.freq	D	ChiSq	Pvalue
269911	0.824	0.00943	2.070	0.15000
1671152	0.832	0.01830	8.430	0.00369
2124090	0.882	-0.00167	0.128	0.72100

Control

	A.freq	D	ChiSq	Pvalue
269911	0.836	0.00782	0.842	0.359
1671152	0.807	0.01290	1.780	0.182
2124090	0.896	0.00458	0.622	0.430

Summary Statistics: Linkage Disequilibrium

PHASE Haplotype Frequencies

	H.freq	H.relfreq
AGA	113	0.115
AGC	534	0.542
ATA	1	0.001
ATC	164	0.166
TGA	2	0.002
TGC	171	0.173
TTC	1	0.001

Linkage Disequilibrium Between Markers

r^2

	269911	1671152	2124090
269911	1.0000	0.0405	0.0233
1671152	0.0405	1.0000	0.0243
2124090	0.0233	0.0243	1.0000

D'

	269911	1671152	2124090
269911	1.000	0.207	0.193
1671152	0.207	1.000	0.192
2124090	0.193	0.192	1.000

P-value

	269911	1671152	2124090
269911	1.00e+00	2.67e-10	1.68e-06
1671152	2.67e-10	1.00e+00	9.84e-07
2124090	1.68e-06	9.84e-07	1.00e+00

Haplotype by GYNGroup

PHASE Haplotypes (All)

	Case	Case(%)	se.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
ATC	66	6.72	1.80	98	9.98	1.62	0.6499	-0.4309
AGC	253	25.76	0.00	281	28.62	0.00	0.8658	-0.1441
TGC	87	8.86	0.42	84	8.55	0.38	1.0392	0.0385
AGA	60	6.11	0.76	53	5.40	0.68	1.1407	0.1316

Pearson Chi-squared Test = 5.6672, DF = 3, P-value = 0.129

Permutation Test P-value = 0.07

haplo.score Haplotypes

	Hap.Freq	Score	P. X^2	P.Sim
ATC	0.1647	-1.8903	0.0587	0.0599
AGC	0.5464	-0.0089	0.9929	0.9960
TGC	0.1695	0.8924	0.3722	0.3749
AGA	0.1104	1.2311	0.2183	0.2106
TGA	0.0053	2.2174	0.0266	0.0105

Global Score = 10.7498, DF = 5, Global P.X^2 = 0.0566, Global P.Sim = 0.0425

Example 11

Haplotype analysis of the *CHGB* locus

[0296] rs454422 and rs236108 are significantly associated at the allele and genotype levels ($P < 0.05$). Moderate LD is observed between 454422 and 236108 ($r^2 = 0.474$). Chi-squared statistics indicate that haplotypes are significantly associated with breast cancer. Haplotype-specific chi-squared values, odds ratios, and score tests indicate that the TAT haplotype contributes most to this effect, suggesting that individuals who carry this haplotype are at a slightly lower risk of developing breast cancer than individuals with other haplotypes.

Statistics

[0297] Chi-squared statistics are estimated to assess whether 1) alleles and genotypes are associated with breast cancer status and 2) marker genotype frequencies deviate significantly from Hardy-Weinberg equilibrium (HWE). Haplotype frequencies and relative frequencies are estimated, as well as several statistics (r^2, D', and p-value) that gauge the extent and stability of linkage disequilibrium between markers in each region. Chi-squared statistics and score tests are estimated to determine whether reconstructed haplotypes are significantly associated with breast cancer status ($P < 0.05$). P-values are estimated for 1) the full set of reconstructed haplotypes and 2) a reduced set that excludes haplotypes with observed frequencies less than 10. Results are presented by chromosome order.

Results

Summary Statistics: Alleles and Genotypes

SNP Locations

SNP.ID	Type	Location
236116	Proximal	5884335
454422	Incident	5891693
236108	Proximal	5916559

Allele by GYNGroup

	N	Case (N=508)	Control (N=536)	Test Statistic
236116 : T	1028	79% (392)	79% (420)	Chi-square=0.08 d.f.=1 P=0.783
454422 : C	998	83% (402)	76% (393)	Chi-square=8.06 d.f.=1 P=0.00452
236108 : T	1016	8% (41)	13% (70)	Chi-square=6.14 d.f.=1 P=0.0132

Genotype by GYNGroup

	N	Case (N=254)	Control (N=268)	Test Statistics
236116 : CC	514	4% (9)	3% (9)	Chi-square=0.22 d.f.=2 P=0.894
CT		34% (84)	36% (96)	
TT		62% (154)	61% (162)	
454422 : AA	499	3% (8)	5% (14)	Chi-square=8.37 d.f.=2 P=0.0152
AC		27% (64)	37% (95)	
CC		70% (169)	58% (149)	
236108 : CC	508	84% (206)	75% (198)	Chi-square=7.04 d.f.=2 P=0.0296
CT		14% (35)	23% (62)	
TT		1% (3)	2% (4)	

Genotype QC: Test of Hardy-Weinberg Proportions

All

	A.freq	D	ChiSq	Pvalue
236116	0.790	-0.00948	1.620	0.203
454422	0.798	0.00356	0.240	0.624
236108	0.893	0.00266	0.381	0.537

Control

	A.freq	D	ChiSq	Pvalue
236116	0.789	-0.01320	1.6200	0.203
454422	0.762	-0.00209	0.0340	0.854
236108	0.869	-0.00150	0.0445	0.833

Summary Statistics: Linkage Disequilibrium

PHASE Haplotype Frequencies

	H.freq	H.relfreq
CCC	207	0.210
TAC	94	0.095
TAT	106	0.107
TCC	581	0.588

Linkage Disequilibrium Between Markers

r^2

	236116	454422	236108
236116	1.0000	0.0673	0.0319
454422	0.0673	1.0000	0.4740

236108	0.0319	0.4740	1.0000

D'

	236116	454422	236108
236116	1.000	0.265	0.265
454422	0.265	1.000	1.000
236108	0.265	1.000	1.000

D' P-value

	236116	454422	236108
236116	1.00e+00	3.33e-16	2.02e-08
454422	3.33e-16	1.00e+00	0.00e+00
236108	2.02e-08	0.00e+00	1.00e+00

Haplotype by GYNGroup

PHASE Haplotypes (All)

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
TAT	39	3.95	2.85	67	6.78	2.65	0.5649	-0.5711
TAC	39	3.95	0.87	55	5.57	0.81	0.6971	-0.3608
CCC	99	10.02	0.01	108	10.93	0.00	0.9074	-0.0972
TCC	299	30.26	1.30	282	28.54	1.21	1.0864	0.0829

Pearson Chi-squared Test = 9.7095, DF = 3, P-value = 0.02120

PHASE Haplotypes (Low Frequency Excluded)

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
TAT	39	4.36	3.17	67	7.49	3.03	0.5630	-0.5745
CCC	99	11.07	0.05	108	12.08	0.05	0.9063	-0.0984
TCC	299	33.45	0.79	282	31.54	0.76	1.0906	0.0867

Pearson Chi-squared Test = 7.8414, DF = 2, P-value = 0.01983

haplo.score Haplotypes

	Hap.Freq	Score	P. X^2	P.Sim
TAT	0.1073	-2.4492	0.0143	0.0130
TAC	0.0951	-1.3720	0.1701	0.1880
CCC	0.2095	-0.1174	0.9065	0.9372
TCC	0.5881	2.4468	0.0144	0.0145

[0298] Modifications may be made to the foregoing without departing from the basic aspects of the invention. Although the invention has been described in substantial detail with reference to one or more specific embodiments, those of skill in the art will recognize that changes may be made to the embodiments specifically disclosed in this application, yet these modifications and improvements are within the scope and spirit of the invention, as set forth in the claims which follow. All publications or patent documents cited in this specification are incorporated herein by reference as if each such publication or document was specifically and individually indicated to be incorporated herein by reference.

[0299] Citation of the above publications or documents is not intended as an admission that any of the foregoing is pertinent prior art, nor does it constitute any admission as to the contents or date of these publications or documents. U.S. patents, documents and other publications referenced herein are hereby incorporated by reference.